Performance Analysis of Different Classifier for Diabetes Diagnosis

Gilliar Meng, Heba Saddeh
Tong College of Information Technology, Egypt

Abstract:

Disease diagnosis is one of the important areas for research. In the last few decades, several computational techniques have been proposed and used for diagnosis of different diseases. In this manuscript, we have tried to compare the performance of different classifiers for early diagnosis of diabetes. The analysis has been carried out over PIMA dataset.

Keywords: Diabetes, Naïve bayes, J48, Classifiers, Data Mining.

1. Introduction:

Diabetes is one of the chronic and lifelong disease in which the human body unable to regulate sugar in blood. The organ pancreas releases the hormone called insulin that helps to convert glucose into energy from the blood[1]. When the body does not make enough insulin it leads to high level of glucose in the blood. Some of the important categories of diabetes are Type1 diabetes, Type2 diabetes, Gestational diabetes (occurs in women during second half of pregnancy and resolved after the delivery of baby), Metabolic syndrome (occur due to high blood pressure and high fat level in blood) and Pre diabetes (a condition in which blood sugar level is higher than normal but not high enough to be considered diabetic). Some of major symptoms associated with diabetic patients are[2][3]:

- increased thirst,
- increased urination,
- increased hunger,
- fatigue,
- blurred vision,
- numbness in the feet or hands,
- wounds that do not heal early
- and unexplained weight loss.

It was observed that about 31.7 million people in India suffered from diabetes[4]. Doctors use some common laboratory tests to diagnose diabetes and its type viz. Finger stick blood glucose, Fasting plasma glucose, Oral glucose tolerance test and Glycosylated haemoglobin test. Dataset used in diabetes are: age, N_preg, PGC, OGTT, DBP, skinthik, insulin, BMI and DPF [13].

The main objective of this manuscript is compare and contrasts the performance of different classifier in exploring PIMA dataset. We have computed different performance metrics. The confusion matrix has been computed and analyzed.

2. Methodology

In the last few years, various researchers have used different computational techniques for diagnosis of different disease among human beings. Some of the important techniques that have been used in different diseases are:

- naïve bayes
- decision tree
- J48
- Decision table
- Support vector machine
- Ensemble based method
- Genetic algorithm
- Firefly algorithm
- Ant colony method
- Simulated annealing etc.

From the past research, we have found that several researchers have used these data mining classifiers for different application like agriculture[5][6], banking[7][8][9], healthcare[10][11][12][13][14][15][[16], sentiment analysis[17][18][19], and education [20][21] etc. In this manuscript, we have considered determined the performance of four major classifiers in examining the PIMA dataset. Different performance metrics like TP rate, FP rate, recall, precision, F-measure, ROC area, root mean squared error, mean absolute error etc have been computed and examined. The basic details of the PIMA dataset are mentioned in Table 1.

Table 1: PIMA dataset

PIMA Dataset	
Instances	768
Attributes	09
List of attributes	Preg, plas, pres, skin, insu, mass, pedi, age, class
Method used	10-fold cross-validation

3. Results and Discussions

Different classifiers have been used for categorization of diabetic patient. The value of different performance metrics obtained when the data have been classified using naïve bayes are given below:

Correctly Classified Instances	586	76.3021 %
Incorrectly Classified Instances	182	23.6979 %
Kappa statistic	0.4664	
Mean absolute error	0.2841	
Root mean squared error	0.4168	
Relative absolute error	62.5028 %	
Root relative squared error	87.4349 %	
Coverage of cases (0.95 level)	97.2656 %	
Mean rel. region size (0.95 level)	83.7891 %	
Total Number of Instances	768	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.844	0.388	0.802	0.844	0.823	0.819	tested_negative
	0.612	0.156	0.678	0.612	0.643	0.819	tested_positive
Weighted Avg.	0.763	0.307	0.759	0.763	0.76	0.819	

The confusion matrix for the same has been mentioned below:

a	b	<-- classified as
422	78	a = tested_negative
104	164	b = tested_positive

The value of different performance metrics obtained when the data have been classified using J48 are given below:

Correctly Classified Instances	567	73.8281 %
Incorrectly Classified Instances	201	26.1719 %
Kappa statistic	0.4164	
Mean absolute error	0.3158	
Root mean squared error	0.4463	
Relative absolute error	69.4841 %	
Root relative squared error	93.6293 %	
Coverage of cases (0.95 level)	95.5729 %	
Mean rel. region size (0.95 level)	89.0625 %	
Total Number of Instances	768	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.814	0.403	0.79	0.814	0.802	0.751	tested_negative
	0.597	0.186	0.632	0.597	0.614	0.751	tested_positive
Weighted Avg.	0.738	0.327	0.735	0.738	0.736	0.751	

The confusion matrix of negative and positive tested cases obtained using J48 classifier is given below:

a	b	<-- classified as
407	93	a = tested_negative
108	160	b = tested_positive

Additionally, a random forest of 10 trees, each constructed while considering 4 random features has been implemented. The value of out of bag error is 0.2747.

Time taken to build model: 0.16 seconds

Correctly Classified Instances	562	73.1771 %
Incorrectly Classified Instances	206	26.8229 %
Kappa statistic	0.3874	
Mean absolute error	0.3128	
Root mean squared error	0.4269	
Relative absolute error	68.8132 %	
Root relative squared error	89.5628 %	
Coverage of cases (0.95 level)	97.3958 %	
Mean rel. region size (0.95 level)	86.4583 %	
Total Number of Instances	768	

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.836	0.463	0.771	0.836	0.802	0.79	tested_negative
0.537	0.164	0.637	0.537	0.583	0.79	tested_positive
Weighted Avg. 0.732	0.358	0.724	0.732	0.726	0.79	

Confusion Matrix

a	b	<-- classified as
418	82	a = tested_negative
124	144	b = tested_positive

The remaining part of this section presents the results obtained using bagging classifier.

Correctly Classified Instances	584	76.0417 %
Incorrectly Classified Instances	184	23.9583 %
Kappa statistic	0.4558	
Mean absolute error	0.311	
Root mean squared error	0.3994	
Relative absolute error	68.4323 %	
Root relative squared error	83.7862 %	
Coverage of cases (0.95 level)	99.8698 %	
Mean rel. region size (0.95 level)	94.5313 %	

Total Number of Instances 768

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.852	0.41	0.795	0.852	0.822	0.829	tested_negative
0.59	0.148	0.681	0.59	0.632	0.829	tested_positive
Weighted Avg. 0.76	0.319	0.755	0.76	0.756	0.829	

Confusion Matrix

a	b	<-- classified as
426	74	a = tested_negative
110	158	b = tested_positive

4. Conclusion

Diabetes is one of the chronic and lifelong disease in which the human body unable to regulate sugar in blood. The organ pancreas releases the hormone called insulin that helps to convert glucose into energy from the blood. In this manuscript, different classifiers like naïve bayes, J48, random forest and bagging have been used to classify the instances of PIMA database. Different metrics like correctly and incorrectly classified instance, kappa statistic, mean absolute error, root mean square error, relative absolute error, root relative squared error, Coverage of cases along with Mean rel. region size (0.95 level) have been computed for these four classifiers. The best performance for correctly classified instances has been achieved with naïve bayes. The rate of correctly classified instance obtained using naïve bayes is 76.3% i.e. out 768 instance, 586 instances were correctly classified by using naïve bayes classifier.

5. References

1. American Diabetes Association. "2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019." Diabetes Care 42.Supplement 1 (2019): S13-S28.
2. American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 33.Supplement 1 (2010): S62-S69.
3. Kaur, Prableen, and Manik Sharma. "Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review." Int. J. Pharm. Sci. Res 9 (2018): 2700-2719.
4. Kaveeshwar, Seema Abhijeet, and Jon Cornwall. "The current state of diabetes mellitus in India." The Australasian medical journal 7.1 (2014): 45.
5. Kale, Shivani S., and Preeti S. Patil. "Data Mining Technology with Fuzzy Logic, Neural Networks and Machine Learning for Agriculture." Data Management, Analytics and Innovation. Springer, Singapore, 2019. 79-87.
6. Nath, Suraj, et al. "Design of Intelligent System in Agriculture using Data Mining." International Journal of Computational Intelligence & IoT 2.3 (2019).
7. Sharma, Manik, Samriti Sharma, and Gurvinder Singh. "Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining." Data 3.4 (2018): 54.
8. Nguyen, Cuong. "The credit risk evaluation models: an application of data mining techniques." (2019).

9. Gulsoy, Nihan, and Sinem Kulluk. "A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.3 (2019): e1299.
10. Bai, BG Mamatha, B. M. Nalini, and Jharna Majumdar. "Analysis and Detection of Diabetes Using Data Mining Techniques—A Big Data Application in Health Care."Emerging Research in Computing, Information, Communication and Applications. Springer, Singapore, 2019. 443-455.
11. Kaur, Prableen, and Manik Sharma. "Diagnosis of Human Psychological Disorders using Supervised Learning and Nature-Inspired Computing Techniques: A Meta-Analysis."Journal of medical systems 43.7 (2019): 204.
12. Rado, Omesaad, et al. "Performance Analysis of Feature Selection Methods for Classification of Healthcare Datasets."Intelligent Computing-Proceedings of the Computing Conference. Springer, Cham, 2019.
13. Sharma, M., G. Singh, and R. Singh. "Stark assessment of lifestyle based human disorders using data mining based learning techniques." IRBM 38.6 (2017): 305-324.
14. Gautam, Ritu, Prableen Kaur, and Manik Sharma. "A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings."Progress in Artificial Intelligence (2019): 1-24.
15. Sharma, Manik, Gurvinder Singh, and Rajinder Singh. "An Advanced Conceptual Diagnostic Healthcare Framework for Diabetes and Cardiovascular Disorders." arXiv preprint arXiv:1901.10530 (2019).
16. Kaur, Prableen, and Manik Sharma. "Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review." Int. J. Pharm. Sci. Res 9 (2018): 2700-2719.
17. Saura, Jose Ramon, Pedro Palos-Sanchez, and Antonio Grilo. "Detecting indicators for startup business success: Sentiment analysis using text data mining." Sustainability 11.3 (2019): 917.
18. Sharma, Manik, Gurvinder Singh, and Rajinder Singh. "Design of GA and Ontology based NLP Frameworks for Online Opinion Mining." Recent Patents on Engineering 13.2 (2019): 159-165.
19. Abbas, Muhammad, et al. "Multinomial Naive Bayes Classification Model for Sentiment Analysis." IJCSNS 19.3 (2019): 62.
20. Tasnim, Nafisa, Mahit Kumar Paul, and AHM Sarowar Sattar. "Identification of Drop Out Students Using Educational Data Mining." 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019.
21. Goel, Pallavi M. "Comparison of Classification Techniques on Data Mining." (2019).