Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection

Conference Paper · November 2016

DOI: 10.1201/9781315364094-82

CITATIONS
69

READS
8,223

4 authors, including:

Dilip Kumar Choubey
Indian Institute of Information Technology Bhagalpur
73 PUBLICATIONS **699** CITATIONS
SEE PROFILE

Sanchita Paul
Birla Institute of Technology, Mesra
52 PUBLICATIONS **876** CITATIONS
SEE PROFILE

Santosh Kumar
Siksha O Anusandhan University
29 PUBLICATIONS **313** CITATIONS
SEE PROFILE

Communication and Computing Systems – Prasad et al. (Eds)
© 2017 Taylor & Francis Group, London, ISBN 978-1-138-02952-1

Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection

Dilip Kumar Choubey, Sanchita Paul & Santosh Kumar
CSE, BIT, Mesra, Ranchi, India

Shankar Kumar
Polytechnic, BIT, Mesra, Ranchi, India

ABSTRACT: Diabetes means blood sugar is above desired level on a sustained basis. The prime objective of this research work is to provide a better classification of diabetes. There are already several existing method, which have been implemented for the classification of diabetes dataset. In medical sector, the classifications systems have been widely used to exploit the patient's data and make the predictive models or build set of rules. In this manuscript firstly NBs used for the classification on all the attributes and then GA used as an attribute selection and NBs used on that selected attribute for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis.

1 INTRODUCTION

Diabetes is a problem and a major public health challenge worldwide. This is one of the most widespread disease, now a day's very common. In this manuscript, Genetic Algorithm (GA) has been used as an attribute (feature) selection method by which four attributes have been selected from eight attributes. Naive Bayes (NBs) are statistical, supervised learning method for classification. Here, NBs has been used for the classification of the diabetes diagnosis.

The paper is organized as follows: Proposed methodology is discussed in section 2, Results and Discussion are devoted to section 3, Conclusion and Future Direction are discussed in section 4.

2 PROPOSED METHODOLOGY

Here, the proposed methodology is implemented by GA as an Attribute Selection and NBs for Classification on PIDD which has been taken from UCI machine learning repository.

The block diagram of proposed approach is shown above and next proposed approach is as follows:

1. The PIDD has been taken from UCI machine learning repository.
2. Apply GA as an Attribute Selection on PIDD.
3. Do the Classification by using NBs on selected attributes and all the attributes in PIDD.

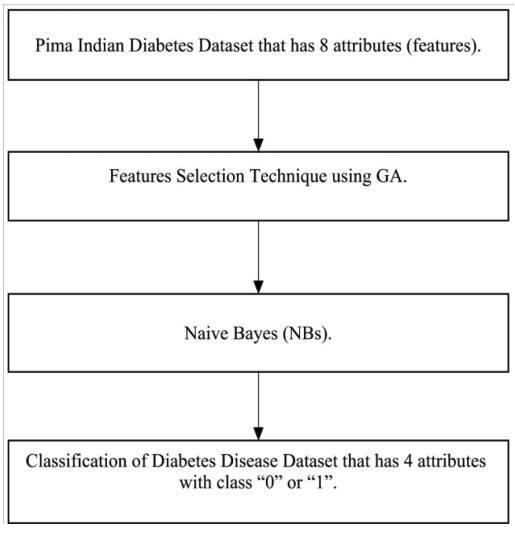

Figure 1. Proposed system.

2.1 *Used diabetes disease dataset*

The Pima Indian Diabetes Dataset (PIDD) has been taken from the UCI Machine Learning repository. The same dataset used in the reference (Polat and Gunes 2007; Seera and Lim 2014; Lukka 2011; Gajni and Abadeh 2011; Choubey and Paul 2016; Ephzibah 2011; Choubey and Paul 2015).

2.2 *GA for attribute selection*

Either the algorithms stops forming new iterations when a maximum number of iterations have been formed or a satisfactory fitness value is achieved for the problem.

The standard pseudo code of GA is given in below algorithm:

Algorithm 1
Begin
$q = 0$
Randomly initialize individual members of population P(q)
Evaluate fitness of each individual of population P (q)
while termination condition is not satisfied do
$q = q+1$
selection (of better fit solutions)
crossover (mating between parents to generate off-springs)
mutation (random change in off-springs)
end while
Return best individual in population;

In Algorithm 1, q represents the iteration counter, initialization is done randomly in search space, and corresponding fitness is evaluated based on the function. After that, GA algorithm requires a cycle of three phases: selection, crossover, and mutation which is briefly explained in (Choubey and Paul, 2016).

2.3 *NBs*

The pseudo code of NBs are discussed below. The same pseudo code (Siddique and Hossain, 2013) of NBs are used for predicting Heart disease.

2.3.1 *Pseudo code*

Calculate diagnosis = "yes", diagnosis = "no" probabilities Pyes, Pno from training input
For each test input samples
For each feature
Calculate of category of feature based on categorical
Division
Calculate probabilities of diagnosis = "yes", diagnosis = "no" corresponds to that category P(feat, yes), P(feat, no) from training input
For each feature
Calculate the resultyes = P(feat, yes), resultno = P(feat, no);
Calculate the resultyes = resultyes *Pyes, resultno = resultno*Pno;
If (resultyes > resultno) then diagnosis = "yes", else then diagnosis = "no".
Where,
Pyes = Total number of yes/Total number of samples or instances;
Pno = Total number of no/Total number of samples;
P(attr, yes) = Total number of yes in corresponding category/Total number of yes;
P(attr, no) = Total number of no in corresponding category/Total number of no;

3 RESULTS AND DISCUSSION

In Experimental studies the dataset have been partitioned between 70–30% (538–230) for training & test of NBs, GA_NBs. It has been performed on PIDD and the results compared with several existing method which is noted in Table 5.

It may be seen that in Table 2, by applying the GA method, four attributes have been selected from eight attributes. This means the cost have reduced to s(x) = 4/8 = 0.5 from 1 and an improvement on the training and classification by a factor of 2.

It is well known that diagnostic performance is usually evaluated in terms of Accuracy, Precision, Recall, Fallout and F–Measure, ROC, Confusion Matrix, Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) which is shown below.

The time taken to build model training set evaluation = 1.22 seconds, and time taken to build model testing set evaluation = 1.14 seconds for NBs Performance. The Table 1 shows the results of both the training set and testing set evaluation by using NBs method for PIDD based on some parameters, which are noted below:

The Table 2 shows the Attribute selection by using GA on PIDD, which is noted below.

Table 1. Results of NBs performance for PIDD.

Measure	Training set evaluation	Testing set evaluation
Precision	0.769	0.766
Recall	0.773	0.77
F–Measure	0.769	0.767
Accuracy	77.3234%	76.9565%
ROC	0.816	0.846
Kappa statistics	0.4875	0.478
Mean Absolute Error	0.2868	0.2768
Root Mean–Squared Error	0.4157	0.3973
Relative Absolute Error	62.9039%	61.0206%
Root Relative Squared Error	87.075%	83.654%

Cofusion Matrix for Training set

a	b	<--classified as
300	49 |	a = tested_ negative
73	116 |	b = tested_ positive

Cofusion Matrix for Testing set.

a	b	<--classified as
128	23 |	a = tested_ negative
30	49 |	b = tested_ positive

Table 2. GA for attributes selection.

Data set	No. of attributes	Name of attributes	No. of instances	No. of classes
PIDD (Without GA)	8	1. Number of times pregnant 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 3. Diastolic blood pressure 4. Triceps skin fold thickness 5. 2 – hour serum insulin 6. Body mass index 7. Diabetes pedigree function 8. Age (years)	768	2
PIDD (With GA)	4	2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 5. 2 – hour serum insulin 6. Body mass index 8. Age (years)	768	2

The time taken to build model training set evaluation = 1.18 seconds, and time taken to build model testing set evaluation = 1.09 seconds for GA_NBs methodology. The Table 3 shows the results of both the training set and testing set evaluation by using NBs method for PIDD on the selected attributes by using GA based on some parameters, which is noted below.

The above figure is the ROC graph for tested_positive class by using GA_NBs on PIDD, achieved 0.844 ROC. It generates very less error rate, and ratio between FPR vs TPR is also good.

The Table 4 shows the analysis of comparison result with and without GA on NBs for PIDD by several measures along with several methods i.e., noted in table.

In the Table 4 it may be seen that with GA the improvement has occurred in every measure except ROC in the case of NBs. The only ROC measure achieved slightly very less, may be by applying this method only on this particular dataset but mostly in any cases by applying attribute selection method improvement occur in every measure.

In the above table, mentioned methods i.e. J48 graft DT, GA_J48 graft DT, MLP NN, GA_MLP NN implemented by Dilip Kumar Choubey et al. have mentioned Precision, Recall, F–Measure, Accuracy, ROC value results in the publication not the Kappa statistics, MAE, RMSE, RAE, RRSE. So once again went to implement the above-mentioned methods to find the not available value results i.e. Kappa statistics, MAE, RMSE, RAE, and RRSE.

Table 3. Results of GA_NBs for PIDD.

Measure	Training set evaluation	Testing set evaluation
Precision	0.75	0.782
Recall	0.757	0.787
F–Measure	0.748	0.78
Accuracy	75.6506%	78.6957%
ROC	0.811	0.844
Kappa statistics	0.4364	0.5021
Mean Absolute Error	0.3105	0.295
Root Mean–Squared Error	0.4158	0.3919
Relative Absolute Error	68.0976%	65.0317%
Root Relative Squared Error	87.1025%	82.5241%

Cofusion Matrix for Training set

a	b	<--classified as
305	44 |	a = tested_ negative
87	102 |	b = tested_ positive

Cofusion Matrix for Testing set

a	b	<--classified as
135	16 |	a = tested_ negative
33	46 |	b = tested_ positive

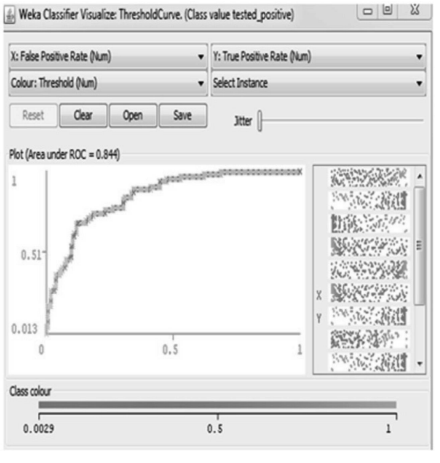

Figure 2. ROC graph for tested_positive class by using GA_NBs methodology on PIDD.

Table 4. Evaluation of NBs & GA _ NBs, along with several method performance for PIDD.

Measure	J48 graft DT Dilip Kumar Choubey et al. (2015)	GA_J48 graft DT Dilip Kumar Choubey et al. (2015)	MLP NN Dilip Kumar Choubey et al. (2016)	GA_MLP NN Dilip Kumar Choubey et al. (2016)	NBs	GA _ NBs
Precision	0.761	0.789	0.781	0.79	0.766	0.782
Recall	0.765	0.748	0.783	0.791	0.77	0.787
F–Measure	0.762	0.754	0.77	0.78	0.767	0.78
Accuracy	76.5217% (0.765217)	74.7826% (0.747826)	78.2609% (0.782609)	79.1304% (0.791304)	76.9565%	78.6957%
ROC	0.765	0.786	0.853	0.842	0.846	0.844
Kappa statistics	0.4665	0.4901	0.4769	0.5011	0.478	0.5021
MAE	0.3353	0.3117	0.2716	0.2984	0.2768	0.295
RMSE	0.4292	0.4114	0.387	0.387	0.3973	0.3919
RAE	73.9186%	68.7038%	59.8716%	65.7734%	61.0206%	65.0317%
RRSE	90.3686%	86.6146%	81.4912%	81.4774%	83.654%	82.5241%

Table 5. Results and comparison with other methods for the PIDD.

Source	Method	Accuracy	ROC
Pasi Luukka (2011)	Sim	75.29%	0.762
	Sim + F1	75.84%	0.703
	Sim + F2	75.97%	0.667
H. Hasan Orkcu et al. (2011)	Binary– coded GA	74.80%	….
	BP	73.80%	….
	Real–coded GA	77.60%	….
Manjeevan Seera et al. (2014)	FMM	69.28%	0.661
	FMM– CART	71.35%	0.683
	FMM-CART– RF	78.39%	0.732
Dilip Kumar Choubey et al. (2015)	J48 graft DT	76.5217%	0.765
	GA_J48 grft DT	74.7826%	0.786
Dilip Kumar Choubey et al. (2016)	MLP NN	78.2609%	0.853
	GA_MLP NN	79.1304%	0.842
Our Study	NBs	76.9565%	0.846
	GA _NBs	78.6957%	0.844

In the Table 5, It may be seen that there are already several existed method for PIDD. The Table 4 shows the result comparison in terms of accuracy and ROC on PIDD for the diagnosis of diabetes. The proposed method i.e., GA_NBs provides the almost highest accuracy and better ROC from all other existing method.

4 CONCLUSION AND FUTURE DIRECTION

Diabetes is a problem with your body that causes blood sugar levels to rise higher than normal. Diabetes can cause serious health complications including blindness, blood pressure, heart disease, kidney disease and nerve damage, etc. which is hazardous to health. The PIDD obtained from UCI repository of machine learning databases on which NBs, GA_NBs method have been applied. In this manuscript, firstly the classification has been done on PIDD by using NBs, and then using GA for Attributes selection, and there by performed classification on the selected attributes. The proposed method minimizes the computation cost, computation time and maximizes the ROC and classification accuracy than several other existing methods. With GA, the improvement has been occurred in every measure except ROC as we may compare from Table 1 and 3, may be by applying this method only on this dataset achieved little less ROC but mostly in any cases by applying attribute selection method the ROC also improved however the classification accuracy and several measure has been improved.

For the future research work, we suggest to develop an expert system of diabetes, which will

provide good ROC, accuracy and this is possible to achieve only by using different Attribute selection and classification method which, could significantly decrease healthcare costs via early prediction and diagnosis of diabetes. The proposed method can also be used for other kinds of diseases but not sure that in all the medical diseases either same or greater than the existing results.

REFERENCES

Choubey, Dilip Kumar., Paul, Sanchita. (2016) 'GA_ MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis', International Journal of Intelligent Systems and Applications (IJISA), MECS, ISSN: 2074–904X (Print), ISSN: 2074–9058. (Online), Vol. 8, No. 1, pp. 49–59.

Choubey, Dilip Kumar., Paul, Sanchita. (2015) 'GA_ J48 graft DT: A Hybrid Intelligent System for Diabetes Disease Diagnosis', International Journal of Bio-Science and Bio-Technology (IJBSBT), SERSC, ISSN: 2233–7849, Vol. 7, No. 5, pp. 135–150.

Ephzibah, E.P. (2011) 'Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis' International Journal on Soft Computing (IJSC), Vol. 2, No. 1.

Ganji, Mostafa Fathi., Abadeh, Mohammad Saniee. (2011) 'A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis' Expert Systems with Applications, Elsevier, Vol. 38, pp. 14650–14659.

Luukka, Pasi. (2011) 'Feature selection using fuzzy entropy measures with similarity classifier', Expert Systems with Applications, Elsevier, Vol. 38, pp. 4600–4607.

Polat, Kemal., Gunes, Salih. (2007) 'An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease', Digital Signal Processing, Elsevier, Vol. 17, pp. 702–710.

Seera, Manjeevan., Lim, Chee Peng. (2014) 'A hybrid intelligent system for medical data classification', Expert Systems with Applications, Elsevier, Vol. 41 pp. 2239–2249.

Siddique, Aieman Quadir., Hossain, Md. Saddam. (2013) 'Predicting Heart-Disease from Medical Data by Applying Naive Bayes and Apriori Algorithm', International Journal of Scientific and Engineering Research (IJSER), Vol. 4, Issue 10.

UCI Repository of Bioinformatics Databases [online] Available: http://www.ics.uci.edu./~mlearn/ML Repository.html.