IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 5 (May. - Jun. 2013), PP 70-73
www.iosrjournals.org

Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data

Niyati Gupta[1], Arushi Rawal[1], Dr. V.L. Narasimhan[2], Savita Shiwani[1]
[1](Dept. of Computer Science and Engineering, Gyan Vihar University, Jaipur, Rajasthan, India)
[2](Dept. of Computer Science, East Carolina University, Greenville, USA)

Abstract: *Healthcare industry is a type of industry, where the data is very large and sensitive. The data is required to be handled very carefully without any mismanagement. There are various data mining techniques that have been used in healthcare industry but the research that has to be done now is on the performance of the various classification techniques. So that amongst the all, the best on can be chosen. In this paper, we aim to consider the accuracy percentage, sensitivity percentage and specificity percentage to provide a result.*
Keywords: *accuracy, classification, data mining, healthcare industry, sensitivity, specificity.*

I. Introduction

There has been a lot of work that has been already done using data mining in healthcare industry and due to enormous success, the people are getting more and more interested in this field .The dataset chosen by us is the diabetes data set. We chose it because it is a common disease throughout the world. The Pima Indians data set is widely used in the diabetic studies because Pima Indians generally show the symptoms of Type-2 diabetes.

The aim of this study is to find the accuracy, sensitivity and specificity percentage of various classification methods. We here have tried to compare the results of various classification methods in Weka with the classification methods implemented on few other tools like Rapidminer and matlab on the basis of few ROC parameters.

II. Literature Survey

There are various numbers of researchers that have been reported working on diagnosis of common diseases. In a study, the diagnosis of Pima Indian diabetes by using the evolving fuzzy approach, has been addressed by Lekas et al.

Sapna and Tamilaras [1], proposed a research that was based on the concept that diabetic mellitus is a cause for nerve disorder .In another study by Jeatarbul and Wong [2], the classification performance is obtained by various types of neural networks i.e. Back Propagation NN, Radial Basis Function NN, General Regression NN, Probabilistic NN and Complementary NN.

A new SSVM for the problems of classification was proposed by Santi Waulan et al [3]. Radha and Rajagopalan, to diagnosis of diabetes introduced an application of fuzzy logic.[4].

In a study, Huang et al used data mining to detect key determinants of Type-2 diabetes.

III. Description Of The Data Set

The dataset used in our study is of Pima Indians who have symptoms of diabetes. The Pima Indians data set is publically available at UCI [5]. There are 768 observations and 9 attributes with no missing values, but there are few impossible values. All the patients in this data set are females, living near Arizona.
The attributes are shown in the table below:

Attribute no	Attribute
1	Number of times pregnant
2	Plasma glucose concentration
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin(mu U/ml)
6	Body mass index(kg/m^2)
7	Diabetes pedigree function
8	Age
9	Class 0 or 1

Table 1: Attributes

The characteristics of the dataset used are shown below in a summarized format.

Data set	No. of example	Input attributes	Output classes	Total No. of attributes	Missing attributes status	Noisy attributes status
Pima	768	8	2	9	No	No

Table 2: Characteristics of data set

IV. Methodology

Several algorithms with running parameters are explained below:

Multilayer Perceptron (MLP)-It is a neural network classification algorithm and is very commonly used. It is experimented in this study with the parameters: learning rate =0.3/0.15; momentum = 0.2; random speed =0; validation threshold= 20; number of epochs =500[6].

BayesNet- This algorithm considers two assumptions: nominal values and no missing values. For estimating the conditional probability tables of network, there are two different parts. Simple estimator and k2 search algorithm are used to run the bayesnet [7].

J48graft -J48 or J48graft is the weka version of C4.5 classifier. Selected parameters are: confidence factor=0.25; min num obj =2; subtree raising=true; unpruned=false [8].

JRip -In this algorithm to produce error reduction, repeated incremental pruning is done. The implementation of JRip in weka is done with parameters: folds= 10; min no. =2; optimizations=2; seed=1; use pruning =true [9].

The classification methods used on various other tools:

PNN – stands for "Probabilistic Neural Network". As the word Neural is so the working is like feed forward neural network. This algorithm is normally used for the classification. It consists of 4 layers

1. Input Layer
2. Hidden Layer
3. Output Layer

LVQ - Learning Vector Quantization discovers the network architecture that has been used for clustering.

FFN - Feed Froward network is in one direction flow of data directly from input nodes to hidden nodes to output nodes.

CFN - Cascade forward networks works like FFN but there is a dependency between the nodes.

DTDN - Distributed Time delay Networks is a time sensitive FFN kind off network where there is a delay associated with every node thus there is a finite dynamic response to time series input.

TDN - Time delay network has a tapped delay line at the input weight in which the dynamics appear only at the input layer of a static multilayer feed forward network.[1]

GINI - Two separate parts are created in the form of binary divisions, each output is separated according to the calculation –

- $Gini_{left}$

$$: 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{left}|} \right)^2$$

- $Gini_{right}$

$$1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{right}|} \right)^2$$

AIS - , artificial immune systems (AIS) are a class of computationally intelligent systems inspired by the principles and processes of the vertebrate immune system[wiki].

V. Roc Parameters

ROC parameters are used to compare the results of various classifiers. For accuracy percentage, sensitivity percentage and specificity percentage, TP, TN, FP AND FN expressions are used.

The abbreviations of the above mentioned ROC parameters and their explanation is shown in the table given below:

Abbreviation	Exposition
TP - True Positive	The number of people who actually suffer from "diabetes" among those who were diagnosed "diabetic."
TN - True Negative	States the number of people who are "healthy" among those who were diagnosed "diabetic".
FP - False Positive	Depicts the number of persons who are unhealthy i.e. is " diabetic" but were diagnosed as "healthy" .
FN - False Negative	The number of people found to be "healthy" among those who were diagnosed as "diabetic".

Table: 3 the ROC PARAMETERS

Various formulas based on ROC parameters are given below:
ACCURACY = (TP+TN)/ (TP+TN+FP+FN)
SENSITIVITY= (TP)/ (TP+FN)
SPECIFICITY= (TN)/ (TN+FP)

VI. Results Of Study

To compare the various classification methods namely Multilayer Perception(MLP) , Neural Network, Bayes Network Classifier , J48graft (C4.5), JRip, Fuzzy Lattice Reasoning (FLR) we worked on weka and the other algorithms like PNN , LVQ,FFN ,CFN, DTDN,TDN are implemented on MATLAB , while GINI algorithm is implemented on Rapidminer. The accuracy, sensitivity and specificity is determined for each algorithms based on the values of various ROC parameters and the formulas. The results of the classifiers in Weka are shown in table 4, while the results of other algorithms are shown in table 5.

	Accuracy%	Sensitivity%	Specificity%
MLP	79.19 %	61.1%	82.6%
BAYESNET	78.98%	60.8 %	81.6%
J48GRAFT	81.33%	59.7 %	81.4%
JRIP	80.91%	55.5 %	83.8 %

Table: 4 Accuracy, sensitivity and specificity percentages using Weka classifiers.

	accuracy	sensitivity	specificity
PNN	72.0%	63.33%	76.9%
LVQ	73.6%	54.4%	84.4 %
FFN	68.8%	54.4%	76.9%
CFN	68.0%	62.2%	71.3%
DTDN	76.0%	53.3%	88.8%
TDN	66.0%	41.1%	81.3%
GINI	65.0%	44.7%	77.8%
AIS	68.8%	52.2%	78.1%

Table: 5 Accuracy, sensitivity and specificity percentages using classifiers on various other tools.

VII. Result Analysis And Conclusion

As we all are aware that clustering is done to find the clusters in a way that the data that fall under same clusters is similar and dissimilar data lie in other clusters. We have done clustering here to divide the patients in clusters based on low risk and high risk of being sick.

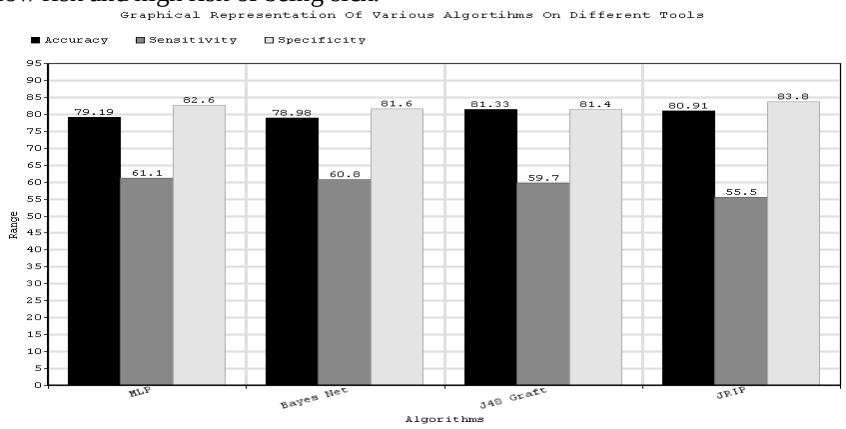

Figure 1

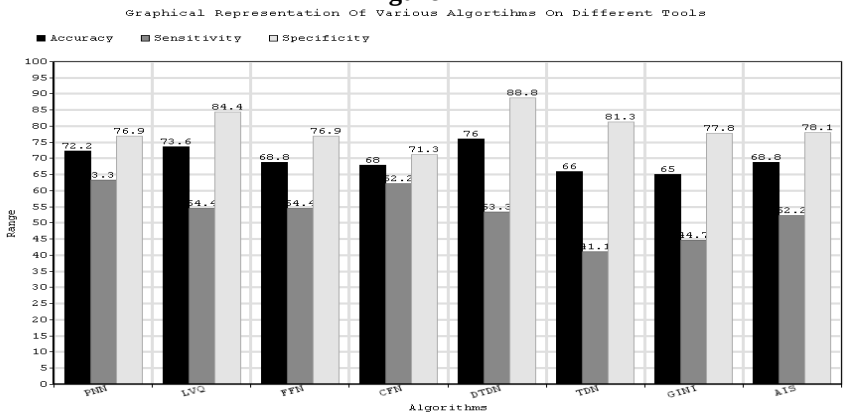

Figure-2

In this study, we have taken various classification methods and compared the results of various algorithms on Weka on the basis of accuracy, sensitivity and specificity with the results of various other algorithms implemented on Matlab and Rapidminer. According to the figure 1 and table 4, we can determine that J48 has the highest accuracy percentage of 81.33 %, while if we consider the other parameters then PNN has the highest sensitivity of 63.33% and DTDN has the highest specificity of 88.8%.

References

[1] M. S. Sapna and D. A. Tamilarasi, "Fuzzy Relational Equation in Preventing Neuropathy Diabetic", *Internati- onal Journal of Recent Trends in Engineering*, Vol. 2, No. 4, 2009, p. 126.

[2] Jeatrakul and K. W. Wong, "Comparing the Perform-ance of Different Neural Networks for Binary Classifica-tion Problems," *The* 8*th International Symposium on Na- tural Language Processing*, Bangkok, 20-22 October 2009, pp. 111-115. doi:10.1109/SNLP.2009.5340935

[3] S. W. Purnami, A. Embong, J. M. Zain and S. P. Rahayu, "A New Smooth Support Vector Machine and Its Appli- cations in Diabetes Disease Diagnosis," *Journal of Com-puter Science*, Vol. 5, No. 12, pp. 1006-1011.

[4] R. Radha and S. P. Rajagopalan, "Fuzzy Logic Approach for Diagnosis of Diabetes," *Information Technology Jour- nal*, Vol. 6, No. 1, pp. 96-102. doi:10.3923/itj.2007.96.102

[5] UCI Machine Learning Repository. http://www.ics.uci.edu/mlearn/MLRepository.html

[6] P. Werbos, "Beyond Regression: New Tools for Predi- ction and Analysis in the Behavioral Sciences," Ph.D. Thesis, Harvard University, Cambridge, 1974

[7] G. H. John and P. Langley, "Estimating Continuous Dis-tributions in Bayesian Classifiers," *Proceedings of the* 11*th Conference on Uncertainty in Artificial Intelligence*, San Francisco, 1995, pp. 338-345.

[8] J. Quinlan, "C4.5: Programs for Machine Learning," Mo- rgan Kaufmann, San Mateo, 1993

[9] I.H. Witten and E. Frank, "Data Mining: Practical Ma- chine Learning Tools and Techniques," 2nd Edition, Mor- gan Kaufmann, San Francisco, 2005.