A Practical Comparison among Neural Networks, Bayesian Networks, and Collaborative Filtering in Classifying Diabetes Mellitus Patients

Bondad, Rachelle G. & Pabico, Jaderick P.

Abstract

Machine Learning is a study of systems that allows learning and prediction based from a data. Implementing Machine Learning for medical purposes is one of its useful and important applications. Diabetes Mellitus is a major health concern worldwide. This paper presents a way to improve data evaluation on Diabetes Mellitus by using different machine learning approaches, namely: Neural Networks, Bayesian Networks, and Collaborative Filtering.

Methodology

A. Implementing ANN, Bayesian Networks and Collaborative Filtering

The authentic patient health records served as data set for the study. This data set was obtained from the Practical Fusion, a free Web-based Electronic Health Records (EHR).

Neural Network is composed of input layer, hidden layer and the output layer which are connected by weights. Ten features served as neurons for the input layer, namely: Gender, Height, Weight, BMI, Systolic, Diastolic, Respiratory Rate, Temperature, Smoking Status and Allergy. The number of neurons in hidden layer was incremented in this study. The output layer consisted of a neuron that served as indicator if patient has Diabetes.

For **Bayesian Networks** implementation, we used Genie, user friendly software for determining graphical decision theoretic models to determine the Direct Acyclic Graph (DAG). Continuous values like Weight, Height, BMI, etc. were discretized first while discrete values like Gender, Allergy and Smoking Status remained the same. Users are free to set the parameters such as Background knowledge, Max Parent Count, Iterations, Seed, Max Time and many more.

Collaborative Filtering is used to make prediction $P(a, i)$ on the active patient *a* (testing) for feature *i*, based on the similarity between patient *a* and other patient *u* who has previously provided a value for that feature. Three collaborative filtering algorithms were implemented, namely: User-based CF using Pearson Correlation, User-Based CF using Vector Cosine Correlation and Item-based CF using Pearson Correlation.

Receiver Operating Characteristic, also known as **ROC Curve**, was used to further analyze the classification done by the algorithms. It is created by plotting the fraction of True Positive Rate (TPR) versus the fraction of False Positive Rate (FPR) for different cut-off points of a parameter.

(a) True Positive Rate or Sensitivity

$$TPR = \frac{TP}{TP + FN}$$

(b) False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

(c) Accuracy

$$ACC = \frac{TP + TN}{P + N}$$

(d) Specificity or True Negative Rate

$$SPC = 1 - FPR$$

Results and Discussion

NEURAL NETWORKS

Two implementations of neural networks were made in this study - manually coded with intended improvements in the basic neural network implementation and using NeuroShell software. The test scores for each ROC measure are as follows:

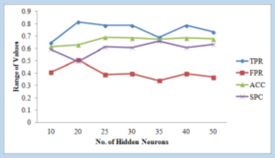

Fig. 1 Results for manually coded ANN

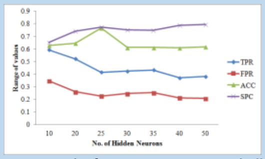

Fig. 2 Results for ANN using NeuroShell

Interpreting results of both methods, they arrived at same result in terms of *Accuracy* rate which led us to NN with 25 Hidden neurons.

BAYESIAN NETWORKS

We can observe that test scores for TPR and FPR of Bayesian Networks are relatively lower compared to the scores in Neural Networks. Lower TPR only means that the network had lower hit rate while low FPR indicates low Type I error. Test score for ACC is not far from the score of Neural Networks.

Measure	Bayesian
TPR	0.38596
FPR	0.23667
ACC	0.60038
SPC	0.76333

Table 1 ROC Results for Bayesian Networks

COLLABORATIVE FILTERING

Based on results, Item-based CF using Pearson correlation had the highest *Sensitivity/*hit rate, while User-based CF had the lowest FPR resulting to a low Type I error. It also showed that the most *accurate* among the three is Item-based using Pearson Correlation.

	Pearson Correlation UB	Pearson Correlation IB	Vector Cosine UB
TPR	0.39912	0.53070	0.43421
FPR	0.29333	0.35333	0.32667
ACC	0.57386	0.59659	0.57008
SPC	0.70667	0.64667	0.67333

Table 2 ROC Results for Collaborative Filtering

ROC curve allows us to analyze and compare all the intelligence methods used in the study. In Fig. 3, all plots lie above the *line of no Discrimination,* so we can say that all methods in the study resulted to a better classification. The best method can be determined by extending the diagonal line upwards and looking at the point that would last touch the line. However, it is unclear in the graph because 25, 30 and 40 Hidden neurons likely occupy same spot.

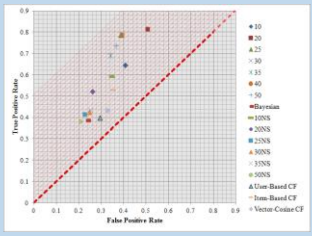

Fig. 3 ROC Space of all algorithms used in study

Another way of finding the best method is by *perfect classification* or by looking at the plot that has the highest sensitivity rate (100% TPR) and the highest specificity rate (100% SPC) or lowest FPR. We therefore determine the best classification method by choosing the plot that has the shortest distance from the *perfect classification point* using distance formula which is shown on table 3.

Methods	Distance
10 Hidden Neurons	0.53915
20 Hidden Neurons	0.53967
25 Hidden Neurons	0.44116
30 Hidden Neurons	0.44701
35 Hidden Neurons	0.45993
40 Hidden Neurons	0.44701
50 Hidden Neurons	0.45269
10 NeuroShell	0.53469
20 NeuroShell	0.54403
25 NeuroShell	0.62651
30 NeuroShell	0.62589
35 NeuroShell	0.62045
40 NeuroShell	0.66355
50 NeuroShell	0.65303
Bayesian Networks	0.65807
User-based CF using Pearson Correlation	0.66865
User-based CF using Vector Cosine	0.65332
Item-based CF using Pearson Correlation	0.58744

Table 3 Distance from Perfect Classification Point

Conclusion

In this problem only, we can say that by using ROC Curve as evaluation method, Neural Networks with 25 Hidden neurons is the best and the most accurate method to use in classifying Diabetes Mellitus patients among Neutral Networks, Bayesian Networks and Collaborative Filtering.

Author

Rachelle G. Bondad is a BSCS student in the Institute of Computer Science, University of the Philippines Los Baños. She is the second child among the three children of Gilda and Francisco Bondad. Her interests are reading ebooks and hanging out with friends.