MSA 2020 AI & Advanced Analytics Project 1

Liang Pan
lpan2573@uni.sydney.edu.au

https://github.com/liangpann/Microsoft-AI-Project-1.git

1. Aim

The primary aim of this project is to investigate the effectiveness of different Machine Learning techniques on prevalent issues in society such as Diabetes Detection. My program is a form of Supervised Learning (more specifically Classification) as new medical data is used to determine whether the patient has diabetes or not.

2. Data

The dataset used in this study is the Pima Indian Diabetes dataset. It contains 768 instances described by 8 numeric attributes (shown below). There are two classes - yes and no. Each entry in the dataset corresponds to a patient's record; the attributes are personal characteristics and test measurements; the class shows if the person shows signs of diabetes or not. There are 500 "no" instances and 268 "yes" instances. The patients are from a Pima Indian heritage, hence the name of the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are female at least 21 years old of Pima Indian heritage. The attributes included are:

1. Number of times pregnant
2. Plasma glucose concentration after 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U / ml)
6. Body mass index (weight in kg / (height in m) ^ 2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable ("yes" or "no")

The original dataset is sourced from the UCI Machine Learning Repository.

3. Results

In order for comparisons to be made, the accuracy of my model was compared to the Weka classifiers shown in the table below:

To evaluate the performance of my classifiers, I implemented 10-fold stratified cross-validation as an extension of my classifier code. My program was able to show the algorithm's average accuracy over the 10 folds. I have included the file pima-folds.csv which contains 10 folds, each containing approximately the same number of examples, and the ratio of yes examples to no examples are approximately the same for each fold.

	1NN	**5NN**	**NB**
Weka Classifier	67.84%	74.48%	75.13%
My Classifier	68.49%	75.52%	75.26%

4. Conclusion

Our results suggest that the number of neighbours in the KNN classifier is linked to the classifier's accuracy. By comparing the results for 1NN and 5NN in both Weka and our own model, we can see that the 5NN classifier has a higher accuracy than 1NN. However, during our implementation, we noticed that 5NN ran slower than 1NN. Considering that we only tested two values of k-Nearest Neighbour, we do not have a definite conclusion on the type of correlation between the number of neighbours and accuracy.

A possible improvement would be to increase the number of attributes in the dataset which would avoid this decrease in accuracy for some classifiers and open up new opportunities for more correlated classifications. Alternatively, introducing feature weighting to be used with CFS could potentially result in better performance. However, more testing is required to make any substantiated claims. This model accurately classifies patients as either having diabetes or not.