Predicting Diabetes in the Pima Indians: An Investigation into Classification Strategies

Group 23: 490424010, 490390494

May 11, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

The aim of this study is to investigate methods for predicting the future onset of diabetes mellitus (or simply diabetes), with relevance to females over 21 and of Pima Indian heritage.

1.2 Relevance

This study is important because the classifiers created have the potential to forewarn individuals of their risk to diabetes, and to be used as an easy clinical tool for early prevention. This is particularly important because, if left untreated, diabetes can lead to many serious long-term health implications such as cardiovascular disease, stroke, diabetic ketoacidosis and even death [1].

2 Data

The dataset used throughout this paper originates from the National Institute of Diabetes and Digestive and Kidney Diseases and was first used in a demonstration of the ADAP Learning Algorithm in 1988 [4]. It consists of 768 non-diabetic females aged at least 21 years old and of Pima Indian heritage. There are 9 columns per row, the first 8 of which are biometric measurement attributes whilst the final one is the class consisting of whether or not the individual with be diagnosed with diabetes. A description of each column in the dataset is shown in Table 1. To maintain consistency the dataset has been cleaned to remove any missing values.

Table 1: A synopsis of the dataset's columns with those selected by CFS highlighted.

Description	**Units**
Number of times pregnant	n/a
Plasma glucose concentration at 2 hours in an oral glucose tolerance test	mg/dL
Diastolic blood pressure	mm Hg
Triceps skin fold thickness	mm
Serum insulin level	μU/mL
Body mass index (BMI)	kg/m^2
Diabetes pedigree function (likelihood of diabetes based on family history)	n/a
Age	years
Is diabetes diagnosed between 1 and 5 years after the above measurements are recorded?	n/a

2.1 Attribute Selection

The Correlation-based Feature Selection (CFS) method is a way of determining a representative set of attributes which are highly correlated with the class but uncorrelated with each other. This can improve the training of a classification model by removing features that are not predictive of the class.

Using the CFS algorithm [3] implemented in Weka 3.8.5 [2], the attributes that were selected are plasma glucose concentration, serum insulin level, BMI, diabetes pedigree function and age, and are additionally highlighted in Table 1.

3 Results and Discussion

3.1 Classifier Accuracy

The canonical Naïve Bayes (NB) and Decision Tree (DT) classification algorithms were implemented with tie decisions resulting in a 'yes' and are hereafter referred to as MyNB and MyDT respectively. 10-fold stratified cross validation was then performed on these algorithms and 12 other inbuilt Weka algorithms using the dataset described in section 2 after normalisation and discretisation for the numeric and nominal classification algorithms respectively.

Tables 2 and 3 present all the resulting accuracy figures for each tested classification algorithm, shown in percentage (%) to 4 d.p., using both the full dataset and the dataset after CFS.

Table 2: The 10-fold stratified cross validation accuracy in percentage (%) of each tested *numeric* classification algorithm using the dataset with and without CFS.

Numeric Data	ZeroR	1R	1NN	5NN	NB	MLP	SVM	MyNB
No feature selection	65.1042	70.8333	67.8385	74.4792	75.1302	75.3906	76.3021	75.2614
CFS	65.1042	70.8333	69.0104	74.4792	76.3021	75.7813	76.6927	76.0407

Table 3: The 10-fold stratified cross validation accuracy in percentage (%) of each tested *nominal* classification algorithm using the dataset with and without CFS.

Nominal Data	DT unpruned	DT pruned	MyDT	Bagg	Boost	RF
No feature selection	75.0000	75.3906	73.4484	74.8698	76.1719	73.1771
CFS	79.4271	79.4271	78.3869	78.5156	78.6458	78.9063

3.2 DT Diagrams

Decision trees were built on the full discretised dataset using three different algorithms: MyDT, and two DT classifiers from Weka (DT unpruned and DT pruned). The two Weka variants were built using J48 (an implementation of the C4.5 algorithm) with default parameters, but differ in that one has been pruned in addition to the other [2]. The DT diagrams are displayed in Figures **??**, 1 and 2.

3.3 Discussion

3.3.1 Comparison of Classifiers

The performance of the classifiers largely varied between different algorithms and datasets.

For the numeric data, there was a large variance in performance between different algorithms, ranging from around 65% to almost 77%.

The best performing algorithm was the SVM, both with and without feature selection, where it achieved an accuracy of 76.9% and 76.3% respectively. Both NB and MLP were similar in performance, generally within only 1% of the SVM accuracy. Therefore this difference may not indicate a significant difference in performance, but could instead be due to random noise in the testing **Validation??** dataset.

On the other hand, the worst performing algorithms were the simplest algorithms, in particular ZeroR, 1R, and 1NN, achieving accuracies between 65% and 71%. These algorithms were likely not complex enough to capture patterns in the data that other, more complex algorithms were able to recognise (i.e. SVM, MLP, NB).

```
glucose = high
| bmi = high
| | triceps = high
| | | npreg = low
| | | | pedigree = high
| | | | | age = high: yes (16.0/5.0)
| | | | | age = low
| | | | | | blood = high: yes (11.0/5.0)
| | | | | | blood = low: no (5.0/2.0)
| | | | pedigree = low
| | | | | blood = high: no (43.0/19.0)
| | | | | blood = low: yes (10.0/4.0)
| | | npreg = high
| | | | blood = high: yes (29.0/8.0)
| | | | blood = low
| | | | | pedigree = high: no (2.0)
| | | | | pedigree = low: yes (3.0)
| | triceps = low: no (13.0/4.0)
| bmi = low: no (29.0/4.0)
glucose = low
| bmi = high
| | insulin = high
| | | age = high
| | | | pedigree = high: yes (7.0/3.0)
| | | | pedigree = low: no (28.0/4.0)
| | | age = low: no (43.0/4.0)
| | insulin = low: no (48.0/2.0)
| bmi = low: no (66.0)
glucose = very high
| insulin = high
| | bmi = high: yes (103.0/16.0)
| | bmi = low
| | | age = high: yes (12.0/3.0)
| | | age = low: no (4.0/1.0)
| insulin = low: no (3.0/1.0)
glucose = medium
| age = high
| | insulin = high
| | | bmi = high
| | | | pedigree = high: yes (37.0/10.0)
| | | | pedigree = low
| | | | | blood = high: no (57.0/24.0)
| | | | | blood = low
| | | | | | triceps = high: yes (15.0/7.0)
| | | | | | triceps = low: no (3.0/1.0)
| | | bmi = low: no (27.0/3.0)
| | insulin = low: no (8.0)
| age = low
| | bmi = high
| | | npreg = low
| | | | triceps = high
| | | | | pedigree = high
| | | | | | blood = high: no (17.0/2.0)
| | | | | | blood = low: yes (7.0/3.0)
| | | | | pedigree = low: no (54.0/8.0)
| | | | triceps = low: no (24.0/1.0)
```

glucose = high
| bmi = high
| | triceps = high: yes (119.0/51.0)
| | triceps = low: no (13.0/4.0)
| bmi = low: no (29.0/4.0)
glucose = low: no (192.0/14.0)
glucose = very high: yes (122.0/24.0)
glucose = medium
| age = high
| | bmi = high
| | | pedigree = high: yes (37.0/10.0)
| | | pedigree = low: no (80.0/33.0)
| | bmi = low: no (30.0/3.0)
| age = low: no (146.0/17.0)

Figure 2: The DT diagram of the Weka J48 algorithm *with* pruning and trained on the full discretised dataset.

Within the nominal data,
however all accuracies were between 65% and 80%.

Nominal better? Weird?? Discussion point? maybe for 3.3.5 or is the data being predicted here actually different? if not its probs just overfitting (to noise) or something and thats something to mention.

3.3.2 Feature Selection

what features did CFS select?

3.3.3 Decision Trees

3.3.4 Tree-based Classifiers

3.3.5 ?Anything else that we consider important

4 Conclusion

conclusion

5 Reflection

References

[1] AE, K., GE, U., JM, M., and JN, F. Hyperglycemic crises in adult patients with diabetes. *Diabetes Care 32*, 7 (Jul 2009), 1335–1343.

[2] Frank, E., Hall, M. A., and Witten, I. H. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4 ed. Morgan Kaufmann, 2016.

[3] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, Apr 1999.

[4] Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care 10* (Nov 1988).

Nomenclature

μU/mL Micro enzyme units per millilitre

CFS Correlation-based feature selection

d.p. decimal points

DT Decision Tree

kg/m^2 Weight in kilograms per height in metres squared

mg/dL Milligrams per decilitre

mm Millimetres

mm Hg Millimetres of mercury

n/a Not applicable

NB Naïve Bayes