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

Group 23: 490424010, 490390494

May 14, 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, and coloured for ease of comparison.

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 MyDT tree was built using the ID3 algorithm (without pruning), which recursively builds a tree based on maximum information gain. 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, 2 and 3 in section 6.

3.3 Discussion

3.3.1 Comparison of Classifiers

Overall, the accuracy of the 14 classifiers ranged roughly between 65% and 80% with a mean of ∼74.5%.

The best performing numeric classifier was the SVM, both with and without feature selection, where it achieved an accuracy of ∼76.7% and ∼76.3% respectively. Similar in performance were MyNB, NB and MLP, with accuracies roughly within 1% of the SVM. This small difference in accuracies ranging from 75% to 77% is not necessarily indicative of algorithmic superiority but may be the effect of random noise in the testing dataset.

On the other hand, the worst performing numeric classifiers were ZeroR, 1R and 1NN, achieving accuracies between 65% and 71%. These simple algorithms are clearly not complex enough to capture patterns in the data, but are instead good points for comparison as to what is easily achievable (for example by predicting the majority class in ZeroR).

Within the nominal classifiers, the highest accuracy was ∼79.4%, and was obtained by both the pruned and unpruned DT using feature selection. Despite this, all of the nominal classifiers performed well using feature selection, with accuracies ranging roughly between 78% and 79.5%. Without feature selection, the best performing nominal classifier was Boost with an accuracy of ∼76.2%.

The worst performing nominal classifier was MyDT, with and without feature selection, where it achieved an accuracy of ∼78.3% and ∼73.4% respectively.

The 6 nominal classifiers clearly performed much better than the 8 numeric ones with a mean accuracy of ∼76.8% compared to ∼72.8%. In addition, using CFS improved or equalled the performance of every classifier, with an average improvement in accuracy of ∼2.1%.

The implementations of MyNB and Weka's NB only differ in terms of their running time performance. In fact, the minimal differences in accuracies evident in Table 2 are most likely the result of different 10-fold data stratifications used in the cross validation accuracy calculations. On the other hand, the implementations of MyDT and Weka's two DTs differ profoundly. MyDT is built using the ID3 algorithm without pruning, whilst Weka uses J48 (an implementation of the 8th revision of the C4.5 algorithm [2]) which is very similar to ID3 but using the normalised information gain ratio as its splitting criterion. This resulted in Weka's two DTs performing better than MyDT with and without feature selection.

3.3.2 Feature Selection

The feature selection method (CFS) selected a subset of five features from the original nine. They were, in no particular order:

- Glucose
- Insulin
- BMI
- Pedigree
- Age

As the aim of CFS is to create a subset of features with high correlation to the class but low correlation to each other, we analyse whether the included and excluded features are as expected.

Features relating to glucose concentration and insulin have been found to be one of the strongest predictors of future diabetes in numerous other studies containing a wide range of predictors [a][b][c][d]. Although a high (negative) correlation between plasma glucose concentration and insulin levels has been demonstrated [e] which should reduce the likelihood of selected both features in CFS, it seems their correlation with diabetes is strong enough to have both features included.

Similarly, BMI and other weight related features have been found to be a good predictor of future diabetes [a][c][d]. Within our dataset, both BMI and Triceps skin fold thickness are weight-related features and have been shown to have significant mutual correlation [f], which could explain why CFS only included one of these features (BMI). Comparatively, BMI has been found to have higher association with health-related risk factors than triceps skin fold thickness [g][h], which may explain why it was prioritised by CFS as a predictor for diabetes.

The other features selected by CFS have also been found to have high correlation with diabetes, which also remaining relatively independent from other risk factors. In particular, age has been show to be both a strong predictor [i] and also relatively uncorrelated with other features, as demonstrated by the inclusion of age in PCA on the same dataset [c]. Family history of diabetes has also been shown to be a relatively strong and independent predictor of diabetes, explaining its inclusion in CFS [j].

[a] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2551654/ this one says best predictors are closely related to glucose and BMI from a selection of 21

[b] https://care.diabetesjournals.org/content/30/6/1544 this one says insulin is significant

[c] https://www.sciencedirect.com/science/article/pii/S2352914819300176 this is literally the same as us. same dataset, they used PCA-based method, which is just for uncorrelation i think. got Glucose, BMI, blood pressure, age

[d] https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6 dataset also had age, blood pressure. selected glucose, bmi with other stuff. other factors might be better, but it at least didn't select blood pressure as expected

[e] https://pubmed.ncbi.nlm.nih.gov/12919921/ negative coreraltion between glucose nad insulin

[f] https://www.ijcmph.com/index.php/ijcmph/article/view/6059 high correlation bmi to skin fold

[g] https://files.eric.ed.gov/fulltext/EJ1201447.pdf [h] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26
both show BMI is more related to heath risk factors than skin fold thickness, even after conjtrolling for a numebr of hter factors. small but statistically significant difference

[i] https://www.sciencedirect.com/science/article/abs/pii/S1871402116303009

[j] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4038917/

Across all models, when training using features selected by CFS the accuracy was always at least as good as no feature selection. This affect was most apparent with tree-based models, with the accuracy increasing 5% in some cases (i.e. DT unpruned, MyDT, RF). This is likely due to a reduction in over fitting, stemming from the removal of features that did not add much information as they were already highly correlated with other features - a key component of CFS. When highly correlated features were included, the algorithms would use this extra feature to fit to noise in the data.

Another advantage of CFS is the reduction in data size with little or no decrease in performance. This can be especially powerful in pruning large datasets that have many features which are highly correlated with each other or poorly correlated with the target classification. In our dataset for example, the number of features was reduced by over 40%. Not only does this allow for smaller file sizes and therefore faster training, but it can also increase the interpretability of models as there are less features being used (e.g. when visualising decision trees).

3.3.3 Decision Trees

- similarity: glucose was used as first split for all trees, second split level is also similar - difference: much larger than equivalent unpruned, also less accurate suggesting overfitting - then segway into generic desc of pruning. how it works, how it leads to shorter tree and still has more accuracy

3.3.4 Tree-based Classifiers

how is this different from overall comparison of classifiers? basically just a comparison of nominal stuff dont wanna overlap too much i guess overall focuses on nominal vs numeric, and looks at best / worst overall

which DT method was used for dagg/boost/rf?

- boosting good even without CFS. try to speculate why. literature? is there a clear link between algos? boosting creates an iterative ensemble(?) of trees that focus on rows that we failed to predict, this is similar to having a number of uncorrelated features/trees QED? and then once CFS is used this advantage goes away - RF bad? if this uses very short trees we can blame this on inability to capture complexity similar to numeric data. - read literature about DFS J48 to figure out why its much better than other algos / MyDT. prob just generically list the "improvments" over ID3 and just go therefore it performs better. - similar to bagging? bagging good bc reduces overfitting therefore good. if using small trees => still not able to fully capture complexity, therefore not as good as full J48 but better than RF. what tree does this use? if it uses full ID3 trees then this doesn't hold since its worse than them.

3.3.5 ?Anything else that we consider important

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

could also talk about why J48 DT is the best. again, look into specifics of J48 and try to justify that it had all advantages of DT without disadvantages (+ advantages that other algos had).

if ur reading this then im already dead. jks im super busy until like 7pm today so ill turn these into actual paras when i get back.

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
1R	One Rule
Bagg	Bagging
Boost	Boosting
CFS	Correlation-based feature selection
d.p.	decimal points
DT	Decision Tree
kg/m^2	Weight in kilograms per height in metres squared
kNN	k-Nearest Neighbours
mg/dL	Milligrams per decilitre
MLP	Multilayer Perceptron
mm	Millimetres
mm Hg	Millimetres of mercury
n/a	Not applicable
NB	Naïve Bayes
RF	Random Forest
SVM	Support Vector Machines
ZeroR	Classifier that always predicts the majority class

6 Appendix

```
glucose = very high
| insulin = high
| | bmi = high
| | | npreg = high
| | | | pedigree = high: yes (16.0/0.0)
| | | | pedigree = low
| | | | | blood = high: yes (12.0/3.0)
| | | | | blood = low: yes (3.0/1.0)
| | | npreg = low
| | | | age = high
| | | | | pedigree = high
| | | | | | triceps = high
| | | | | | | blood = high: yes (10.0/5.0)
| | | | | | | blood = low: yes (1.0/1.0)
| | | | | | triceps = low: yes (1.0/0.0)
| | | | | pedigree = low
| | | | | | blood = high
| | | | | | | triceps = high: yes (14.0/2.0)
| | | | | | | triceps = low: yes (3.0/0.0)
| | | | | | blood = low
| | | | | | | triceps = high: yes (3.0/0.0)
| | | | | | | triceps = low: yes (1.0/1.0)
| | | | age = low
| | | | | pedigree = high: yes (12.0/0.0)
| | | | | pedigree = low
| | | | | | triceps = high
| | | | | | | blood = high: yes (7.0/2.0)
| | | | | | | blood = low: yes (3.0/0.0)
| | | | | | triceps = low
| | | | | | | blood = high: yes (1.0/0.0)
| | | | | | | blood = low: no (1.0/0.0)
| | bmi = low
| | | age = high
| | | | triceps = high
| | | | | npreg = high
| | | | | | pedigree = high
| | | | | | | blood = high: yes (1.0/1.0)
| | | | | | | blood = low: yes (1.0/0.0)
| | | | | | pedigree = low: yes (2.0/0.0)
| | | | | npreg = low
| | | | | | blood = high: yes (1.0/1.0)
| | | | | | blood = low: yes (1.0/1.0)
| | | | triceps = low: yes (3.0/0.0)
| | | age = low
| | | | blood = high
| | | | | triceps = high: no (1.0/0.0)
| | | | | triceps = low: yes (1.0/1.0)
| | | | blood = low: no (1.0/0.0)
| insulin = low
| | pedigree = high: yes (1.0/0.0)
| | pedigree = low: no (2.0/0.0)
```

```
glucose = high
| bmi = high
| | age = high
| | | pedigree = high
| | | | blood = high
| | | | | npreg = high: yes (12.0/2.0)
| | | | | npreg = low
| | | | | | triceps = high
| | | | | | | insulin = high: yes (9.0/4.0)
| | | | | | | insulin = low: yes (1.0/0.0)
| | | | | | triceps = low: yes (1.0/0.0)
| | | | blood = low
| | | | | npreg = high: no (2.0/0.0)
| | | | | npreg = low: yes (1.0/1.0)
| | | pedigree = low
| | | | insulin = high
| | | | | triceps = high
| | | | | | blood = high
| | | | | | | npreg = high: yes (9.0/6.0)
| | | | | | | npreg = low: no (12.0/11.0)
| | | | | | blood = low
| | | | | | | npreg = high: yes (3.0/0.0)
| | | | | | | npreg = low: yes (2.0/1.0)
| | | | | triceps = low
| | | | | | npreg = high: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | blood = high: yes (1.0/1.0)
| | | | | | | blood = low: no (1.0/0.0)
| | | | insulin = low: yes (1.0/0.0)
| | age = low
| | | triceps = high
| | | | pedigree = high
| | | | | blood = high: yes (6.0/5.0)
| | | | | blood = low: no (3.0/2.0)
| | | | pedigree = low
| | | | | blood = high: no (12.0/8.0)
| | | | | blood = low: yes (4.0/3.0)
| | | triceps = low
| | | | blood = high: no (3.0/0.0)
| | | | blood = low
| | | | | insulin = high
| | | | | | pedigree = high: yes (1.0/1.0)
| | | | | | pedigree = low: no (1.0/0.0)
| | | | | insulin = low: no (1.0/0.0)
| bmi = low
| | triceps = high
| | | insulin = high
| | | | pedigree = high: no (5.0/0.0)
| | | | pedigree = low
| | | | | age = high
| | | | | | blood = high: no (5.0/1.0)
| | | | | | blood = low: no (2.0/0.0)
| | | | | age = low
| | | | | | blood = high: no (1.0/0.0)
| | | | | | blood = low: yes (2.0/2.0)
| | | insulin = low
| | | | pedigree = high: yes (1.0/0.0)
| | | | pedigree = low: no (1.0/0.0)
| | triceps = low: no (9.0/0.0)
glucose = medium
| age = high
| | bmi = high
| | | pedigree = high
| | | | npreg = high: yes (13.0/0.0)
| | | | npreg = low
| | | | | triceps = high
| | | | | | blood = high: yes (9.0/7.0)
| | | | | | blood = low: yes (3.0/3.0)
| | | | | triceps = low: yes (2.0/0.0)
| | | pedigree = low
| | | | insulin = high
| | | | | blood = high
| | | | | | npreg = high: no (14.0/12.0)
| | | | | | npreg = low
| | | | | | | triceps = high: no (18.0/11.0)
| | | | | | | triceps = low: yes (1.0/1.0)
| | | | | blood = low
| | | | | | triceps = high
| | | | | | | npreg = high: yes (3.0/3.0)
| | | | | | | npreg = low: yes (5.0/4.0)
| | | | | | triceps = low: no (2.0/1.0)
| | | | insulin = low: no (5.0/0.0)
| | bmi = low
| | | blood = high
| | | | npreg = high: no (13.0/0.0)
| | | | npreg = low
| | | | | pedigree = high: no (3.0/0.0)
| | | | | pedigree = low
| | | | | | triceps = high: no (2.0/1.0)
| | | | | | triceps = low: no (2.0/0.0)
| | | blood = low
| | | | npreg = high: yes (1.0/0.0)
| | | | npreg = low
| | | | | triceps = high: no (5.0/0.0)
| | | | | triceps = low: no (2.0/1.0)
| age = low
| | bmi = high
| | | triceps = high
| | | | npreg = high: yes (1.0/1.0)
| | | | npreg = low
| | | | | pedigree = high
| | | | | | blood = high
| | | | | | | insulin = high: no (12.0/2.0)
| | | | | | | insulin = low: no (3.0/0.0)
| | | | | | blood = low
| | | | | | | insulin = high: yes (3.0/3.0)
| | | | | | | insulin = low: yes (1.0/0.0)
| | | | | pedigree = low
| | | | | | blood = high
| | | | | | | insulin = high: no (20.0/5.0)
| | | | | | | insulin = low: no (3.0/0.0)
| | | | | | blood = low
| | | | | | | insulin = high: no (18.0/2.0)
| | | | | | | insulin = low: no (5.0/1.0)
| | | triceps = low
| | | | pedigree = high
| | | | | blood = high: no (4.0/0.0)
| | | | | blood = low
| | | | | | insulin = high: no (3.0/1.0)
| | | | | | insulin = low: no (2.0/0.0)
| | | | pedigree = low: no (14.0/0.0)
| | bmi = low
| | | pedigree = high
| | | | insulin = high: no (5.0/0.0)
| | | | insulin = low
| | | | | blood = high: no (1.0/0.0)
| | | | | blood = low: yes (1.0/1.0)
| | | pedigree = low: no (34.0/0.0)
glucose = low
| bmi = high
| | insulin = high
| | | age = high
| | | | pedigree = high
| | | | | blood = high
| | | | | | npreg = high
| | | | | | | triceps = high: yes (2.0/1.0)
| | | | | | | triceps = low: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | triceps = high: no (1.0/0.0)
| | | | | | | triceps = low: yes (1.0/0.0)
| | | | | blood = low: yes (1.0/0.0)
| | | | pedigree = low
| | | | | triceps = high
| | | | | | npreg = high
| | | | | | | blood = high: no (8.0/1.0)
| | | | | | | blood = low: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | blood = high: no (11.0/2.0)
| | | | | | | blood = low: no (3.0/1.0)
| | | | | triceps = low: no (1.0/0.0)
| | | age = low
| | | | blood = high: no (18.0/0.0)
| | | | blood = low
| | | | | triceps = high
| | | | | | pedigree = high: no (5.0/1.0)
| | | | | | pedigree = low: no (9.0/3.0)
| | | | | triceps = low: no (7.0/0.0)
| | insulin = low
| | | blood = high
| | | | age = high: no (12.0/0.0)
| | | | age = low
| | | | | triceps = high
| | | | | | pedigree = high: yes (1.0/0.0)
| | | | | | pedigree = low: no (5.0/1.0)
| | | | | triceps = low: no (6.0/0.0)
| | | blood = low: no (23.0/0.0)
| bmi = low: no (66.0/0.0)
```

Figure 1: The DT diagram of MyDT trained on the full discretised dataset.

```
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)
| | | npreg = high: yes (2.0/1.0)
| | bmi = low: no (42.0/1.0)
```

Figure 2: The DT diagram of the Weka J48 algorithm *without* pruning and trained on the full discretised dataset.

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