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

Group 23: 490424010, 490390494

May 10, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

this study is important because it is worth 24% of our grade.

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 [2]. 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.

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

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 implemented in Weka 3.8.5 [1], the attributes that were selected are highlighted in Table 1.

3 Results & Discussion

All results are 10-fold stratified cross validation accuracy figures in percentage (%).

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

Nominal Data	DT unpruned	DT pruned	MyDT	Bagg	Boost	RF
No feature selection	75	75.3906	73.4484	74.8698	76.1719	73.1771
CFS	79.4271	79.4271	78.3869	78.5156	78.6458	78.9063

J48 unpruned tree

---

a = high
| c = high
| | e = high: yes (82.0/31.0)
| | e = low: no (50.0/21.0)
| c = low: no (29.0/4.0)
a = low
| c = high
| | b = high
| | | e = high
| | | | d = high: yes (7.0/3.0)
| | | | d = low: no (28.0/4.0)
| | | e = low: no (43.0/4.0)
| | b = low: no (48.0/2.0)
| c = low: no (66.0)
a = very high
| b = high
| | c = high: yes (103.0/16.0)
| | c = low
| | | e = high: yes (12.0/3.0)
| | | e = low: no (4.0/1.0)
| b = low: no (3.0/1.0)
a = medium
| e = high
| | c = high
| | | d = high: yes (37.0/10.0)
| | | d = low: no (80.0/33.0)
| | c = low: no (30.0/3.0)
| e = low: no (146.0/17.0)
J48 pruned tree

---

a = high
| c = high
| | e = high: yes (82.0/31.0)
| | e = low: no (50.0/21.0)
| c = low: no (29.0/4.0)
a = low: no (192.0/14.0)
a = very high: yes (122.0/24.0)
a = medium

| e = high
| | c = high
| | | d = high: yes (37.0/10.0)
| | | d = low: no (80.0/33.0)
| | c = low: no (30.0/3.0)
| e = low: no (146.0/17.0)

3.1 Feature Selection

3.2 Comparison of Classifiers

4 Conclusion

conclusion

5 Reflection

References

[1] 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.

[2] 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* (11 1988).

Nomenclature

μU/mL Micro enzyme units per millilitre

CFS Correlation-based feature selection

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