Amazing AI/Weka Report

Aim

The aim is to predict whether someone a Pima Indian female has diabetes based on measureable factors. This prediction can be an important step in detecting those without diabetes that have yet to be diagnosed, as well as identifying individuals at higher risk of having diabetes. The Pima Indian people in particular have a very high proportion of individuals with diabetes[1], so detecting and determining factors that may be correlated is useful.

Data

The data contains eight numeric attributes and one nominal class attribute. Each data point corresponds to an female individual of Pima Indian nationality over the age of 21. The attributes contain information on the number of times an individual has been pregnant, their plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin concentration, body mass index, diabetes pedigree function, and age. The nominal class attributes is one of two values, “yes” or “no”, indicating whether the individual has diabetes.

Whilst there are eight attributes per individual, not all attributes will be an indicator of a person having diabetes. To determine whether an attribute is useful or not, correlation-based feature selection (CFS) was used to select only the attributes that were correlated with having diabetes. Using this method, it was found that the number of times an individual has been pregnant, their diastolic blood pressure, and triceps skinfold thickness were not significantly correlated with the individual having diabetes. Consequently, only five of the eight attributes were selected by CFS, specifically plasma glucose concentration, serum insulin concentration, body mass index, diabetes pedigree function, and age. The advantages of using CFS are discussed below.

Results & Discussion

	ZeroR	1R	1NN	5NN	NB
No feature selection	65.1042%	70.8333%	67.8385%	74.4792%	75.1302%
CFS	65.1042%	70.8333%	69.0104%	74.4792%	76.3021%

	DT	MLP	My1NN	My5NN	MyNB
No feature selection	71.8750%	75.3906%			

[1] http://www.ncbi.nlm.nih.gov/pubmed/7988310

CFS	73.3073%	75.7813%			

Classifier Performance

Feature Selection

For every classifier examined, selecting only the significantly correlated features resulted in equal or greater accuracy than using all the features. This may initially seem counter-intuitive, as a the classifiers produced a more accurate result using less data. The reason for this is because the additional data is not significantly correlated with whether an individual has diabetes or not. This results in overfitting, as the classifiers will attempt to find correlations between the these attributes and diabetes even though there is no correlation. As there will be some random variation in these attributes, the classifiers will learn about that instead.

For ‘dummy’ classifiers, such as ZeroR and 1R, the classifiers performed identically. This is because these classifiers either discard all, or all-but-one of the attributes, and so having more attributes will make no difference. For all the other classifiers, except for 5NN, the classifiers performed the better. One explanation for 5NN performing as well with or without feature selection is that in relation to the class, uncorrelated attributes vary randomly, and so can be considered as noise. Selecting 5 data points has a similar effect as averaging, and removes some noise.

The features selected by CFS do have some intuitive meaning to them. As diabetes is a disease associated with insulin deficiency and blood sugar levels, it is fitting that these attributes are correlated with whether an individual has diabetes.

One additional advantage is that the algorithms run faster using CFS. This is because there are less attributes to analyse. For example, with distance calculations in KNN, a model only has to calculate distance in 5 dimensions, rather than 8. Similarly, for the multi-layer perceptron classifier, there are less perceptrons to be trained and so the model will converge faster.

Conclusion

The models can predict with reasonably high level of accuracy...

Reflection