Aim

Our aim is to implement the K-Nearest Neighbour and the Naïve Bayes algorithms in a real dataset and evaluate them using the stratified cross validation method and comparing to the performance of other classifiers on the same dataset using the software Weka. We also aim to evaluate the effect of feature selection, using a particular Correlation-based Feature Selection method (CFS) from Weka, in the performance of all these classifier methods.

Data

The dataset used for training and testing of the algorithm was donated by Vincent Sigillito from the National Institute of Diabetes and Digestive and Kidney Diseases to our supervisors and modified for the purpose of this assignment. The dataset contains eight attributes of data taken from a group of 768 women with at least 21 years old of Pima Indian heritage. Each data also contains the class, saying if the patient has or not diabetes. The eight attributes are:

1. Number of times pregnant;

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test;

3. Diastolic blood pressure (mmHg);

4. Triceps skin fold thickness (mm);

5. 2-Hour serum insulin (mu U/ml);

6. Body mass index (weight/height in kg/m2);

7. Diabetes pedigree function;

8. Age (years).

For testing the results, we used the software Weka to make a Correlation-based Feature Selection (CFS) to take a subset of attributes that best represent the dataset, considering how good the attributes are at predicting the class and how much they correlate with the other attribute. The attributes selected were:

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight/height in kg/m2)

7. Diabetes pedigree function

8. Age (years)

Results and Discussion

Conclusion

Reflection