Research Report

Assignment 2: Predicting Diabetes COMP3308 Introduction to Artificial Intelligence Semester1, 2012



Weilong Ding (309056942)

Ruijun Gu (309017785)

Table of Contents

Research Aim 3

Data Description 3

Data set used 3

Data preparation 3

Attribute selection 4

Results and discussion 4

Results 4

Discussion 4

Conclusion 5

Reflection 5

Compile Instructions 6

# Research Aim

Nowadays along with the perceptible improvement of standard of living worldwide, diabetes has already become a very important healthy issue in certain countries. This research aims to find out the related symptoms of diabetes and the best way to predict diabetes so that people can take protective measure as soon as possible.

# Data Description

## Data set used

The data set we are to use for the research is Pima Indians Diabetes Database provided by National Institutes of Diabetes and Digestive and Kidney Diseases.

All the instances in the data set are female particularly selected from a larger database.

The data set has 8 attributes plus class:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)

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

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

Class 1 is interpreted as “Test Positive for Diabetes”.

The data set has 768 instances with 500 as Class0 and 268 as Class1. Missing values of some attributes exists in the data set. Certain measurement needs to be taken to handle it.

## Data preparation

Since the original data set only includes raw data, a header line needs to be added into the file for later processing and Class are changed from numeric (1 and 0) to nominal (Class 1 and Class 0). The format still obeys comma-separated-value rule.

Because different attributes are measured on different scales and the effect of attributes with smaller scales will be less significant than those with larger scales, all the attributes (except class) need to be normalized between 0 and 1 so that all the attributes can be guaranteed to contribute and no attributes dominate. The normalization is done in Weka.

One thing needs to mention is that we decide not to handle the missing value since in the data set all the missing value is represented as 0, which is totally possible for some attributes such as number of times pregnant and totally impossible for other attributes such as blood pressure. Therefore to avoid confusion we will take 0 as normal value in our training algorithm.

## Attribute selection

Attribute selection method we use in our research is Correlation-based Feature Selection. CFS is a method for selecting a subset of the original attributes. It searches for “best” subset of features where “best “ is defined by a heuristic which takes into consideration two criteria: 1) how good the individual features are at predicting the class and 2) how much they correlate with the other features. Good subsets of features contain features that are highly correlated with the class and uncorrelated with each other. In our research we want to reduce the noise in the data set as much as possible and find out the most direct factors related with diabetes so that the prediction on diabetes can be much more accurate.

In our research we did attribute selection using Weka. The subset of features selected is:

1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2. Body mass index
3. Diabetes pedigree function
4. Age

# Results and discussion

## Results

The table below shows the testing accuracy results (in %, using 10-fold cross validation). All trainings are done in Weka except MyNB. MyNB is the Naïve Bayes algorithm we use in our research.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ZeroR | 1R | 1-NN | 5-NN | DT | MLP | NB | SVM | MyNB |
| No Feature Selection | 65.1042% | 72.7865% | 70.1823% | 73.1771% | 73.8281% | 75.1302% | 76.4021% | 77.3438% | 75.3879% |
| CFS Feature Selection | 65.1042% | 72.7864% | 68.3594% | 73.8281% | 74.8698% | 75.2604% | 77.474% | 76.9531% | 76.8216% |

## Discussion

As we mention earlier CFS reduces the attributes to 4:

1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2. Body mass index
3. Diabetes pedigree function
4. Age

Comparing the accuracy with CFS and the one without CFS, we can see that for ZeroR and 1R algorithm the accuracies almost stay the same. The accuracies of 1-NN and SVM (Supported Vector Machine) decrease about 2% and 0.4% respectively while the ones of rest of the algorithms all increase more or less. Before feature selection SVM has the highest accuracy (77.3438%) while after feature selection Naïve Bayes (Weka’s) becomes the most accurate (77.474%). The accuracy of Naïve Bayes algorithm implemented by our own increases the most, up to 1.5%. Although the rise of accuracies is not much, CFS does have improved the performances of most of the algorithms and fewer features can reduce the training time. Feature selection is also beneficial for algorithms that work better on uncorrelated attributes such NB since these algorithms naturally think all the attributes are independent.

The selected features also intuitively make sense: High plasma glucose concentration may lead to a high fat index. Body mass index can reflect a person’s state of health directly and Diabetes pedigree function can show us the history of genetic disease of a person’s family, which is a very important factor, correlated with diabetes. We don’t see too much point in the selection of Age. Perhaps diabetes is more commonly seen among youth, which needs further study.

Compared with the Weka’s NB, our NB algorithm is less accurate with about 1% difference with no CFS. After feature selection the difference reduces to 0.65%.

We noticed that after training using Naïve Bayes algorithm, mean and standard deviation we got are slightly different from the ones from Weka, which may be the reason why Weka’s NB is more accurate.

# Conclusion

From our research we can see that Naïve Bayes algorithm with CFS is the most accurate way to predict diabetes. As statistical base learning, Naïve Bayes calculate the mean and standard deviation of all attributes of each class during training since our data set is numeric. After training NB uses probability density function (Normal or Gaussian) to calculate probability and pick the class with the maximum probability as the classified result. Although Naïve Bayes algorithm simplifies computation and has a clear semantics for representing using and learning probabilities knowledge such as predicting diabetes, it still has limitations. Since Naïve Bayes algorithm holds independence assumption, correlated attributes will reduce the power of Naïve Bayes. Therefore in the future work, applying attribute selection beforehand can give us a more accurate result.

# Reflection

In this assignment we were asked to implement Naïve Bayes algorithm with 10-fold validation to predict diabetes, which is a very challenging task. By designing Naïve Bayes algorithm by our own we had a deep understanding about how Naïve Bayes works, especially for numeric data. There are also some tricky parts such as calculating the correct mean and standard deviation for all attributes of each class and implement Probability Density Function in code accurately. There was a trivial mistake, which really worth mentioning: When we wrote the code for PDF, we forgot to bracket a denominator, which lead to that the accuracy we got were always less than the normal one by 10% and it took ages for us to find this tiny mistake. After the problem was fixed, everything went smoothly. From this “accident” we have realized that accuracy is really important not only in the algorithm we implement, but also in our implementation. A tiny mistake may hurt a lot.

Besides this, implementing 10-fold validation is also a tricky part since we need to make sure that the data set is stratified evenly and training set and testing set should not overlap. It is a fiddly work but it did bring us a huge feeling of achievement.

# Compile Instructions

A script named “NB\_KFold” is included at the root folder of the zipped files. To run Naïve Bayes with K-fold validation on an input dataset use the following command:

./NB\_KFold <input\_file> <folds>

where <input\_file> is the name of input file which should be put in bin folder and <folds> will be the number of folds you want to validate the dataset. Folds should be greater than 1.