450021372
440517742

COMP3308 Assignment 2: Classification

Prediction of Diabetes in Pima Indian Women

Aim

The aim of this study is to build and evaluate a classifier that utilises the K-Nearest Neighbour and Naïve Bayes algorithm to predict whether a woman of Pima Indian heritage will show signs of diabetes or not based on their personal characteristics and test measurement. This study would aid in the early detection of diabetes in women which is important from a medical standpoint because diabetes brings about a number of complications that are harder to control, treat and manage the later the diagnosis.

Data

The dataset contains 768 records of women of Pima Indian descent. Each record contains 8 attributes based on their personal characteristics and test measurements as well as a 9th class attribute either “yes”(signs of diabetes present) or “no” (signs of diabetes not present).
The 8 attributes are:

- Pregnancies - Number of times pregnant
- Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Blood Pressure - Diastolic blood pressure (mm Hg)
- Skin Thickness - Triceps skin fold thickness (mm)
- Insulin - 2-Hour serum insulin (mu U/ml)
- BMI - Body mass index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function - Diabetes pedigree function
- Age - Age in years

Weka has a built-in Correlation-based Feature Selection(CFS) algorithm that “evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them”[1]. A good subset of features consist of features that are “highly correlated with the class while having low intercorrelation”[1] with each other.
Running the Pima dataset through Weka’s CFS Best First Search algorithm yielded records with the following attributes:

- Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Insulin - 2-Hour serum insulin (mu U/ml)
- BMI - Body mass index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function - Diabetes pedigree function

- Age - Age in years

Results and Discussion

Weka's Classifier

	ZeroR	1R	1NN	3NN	NB	DT	MLP	SVM
No feature selection	65.10%	70.83%	67.84%	72.66%	75.13%	71.88%	75.39%	76.30%
CFS	65.10%	70.83%	69.01%	73.31%	76.30%	73.31%	75.78%	76.69%

ZeroR: The accuracy has no difference between the original selection and the Correlation- based feature selection. Which are both 65.10%
1R: The Accuracy for no feature selection is 70.83% while the accuracy for CFS stays the same
1NN: The accuracy for the original selection is 67.84%, 1.17% lower than than the accuracy in CFS
3NN: There is a 72.66% of accuracy in no feature selection while there is a 73.31% of accuracy in CFS
NB: The accuracy in no feature selection is 75.13% and the accuracy in CFS is 76.3%
DT: There is a 73.31% of accuracy in CFS, 1.43% higher than that in no feature selection, which has 71.88% of accuracy
MLP: The accuracy in no feature selection is 75.39% while in CFS, it has a 75.78% accuracy
SVM: No feature selection has accuracy of 76.30% and CFS has accuracy of 76.69%. SVM has the highest accuracy both in no feature selection and CFS

Our Classifier

	My1NN	My3NN	My5NN	MyNB
No feature selection	68.49%	74.08%	75.52%	75.26%
CFS	68.23%	73.44%	75.00%	75.91%

My1NN: Accuracy in no feature selection reaches to 68.49%, a little bit higher than 68.23% in CFS
My3NN: No feature selection has 74.08% of accuracy and CFS has 73.44% of accuracy
My5NN: No feature selection has 75.52% of accuracy and CFS has 75% of accuracy. It has the highest performance in no feature selection

450021372
440517742

MyNB: There is a 75.26% accuracy in no feature selection while there is a 75.91% accuracy in CFS which is the highest value in CFS

Through Weka's classifier and our classifier, overall, our classifier has better performance in the accuracy. However, CFS of 1NN has better performance in weka than that in our classifier and also Weka has better accuracy in CFS of naive bayes, which is 76.30%, 0.39% higher than our 75.91%

Weka's CFS algorithm identified a subset of 5 attributes from the 8 that it was given. This subset of attributes selected by Weka were highly intuitive to us. Diabetes is a disease that results from too much sugar(glucose) in the blood which is usually regulated by insulin. Diabetes (Type 2) is also common among the elderly and the overweight(high BMI).

CFS was expected to improve both kNN and NB due to its ability to improve the signal to noise ratio however this is not guaranteed and in some cases can degrade an algorithm's performance especially if there was little to no noise present in the first place. CFS was beneficial only to NB and degraded the performance of our kNN classifier. On Weka, CFS improved the performance of its kNN,NB, MLP and SVM classifiers.

CFS slightly degraded the performance of our kNN algorithm. This could be because despite eliminating uncorrelated features, there maybe be outliers present in the dataset contained by the selected subset of features which would cause overfitting as kNN is very sensitive to anomalies. Weka may have performed better as they have tools built in to detect anomalies in datasets.

It was expected of CFS to affect NB because NB assumes that feature values within classes are independent of each other. NB can be affected a lot by noise or redundant attributes. If there are 2 correlated features then treating them as independent means that both those features individually impact the prediction more than they should. Despite this, NB is known to still work well when moderate dependencies exist. This is because the calculated probability is not so far off that it would misclassify.

Conclusion

The study has gave us a more clear view about diabetes in Pima Indian women. Since the accuracy of prediction improved after correlation- based feature selection, The statistics demonstrate an evidence that there is a consistent correlation between the attributes diabetes and glucose, Insulin, BMI, diabetes pedigree function and age that link to diabetes. Keep an eye on the index of these attributes would help them avoid

diabetes at the early stage and people may find the solution of diabetes from these 5 attributes.

It was expected that both of our algorithms kNN and NB would benefit from CFS, however only NB did. In fact, NB benefited from CFS the most across both our classifiers and Weka's. In Weka, CFS improved NB more so than it did kNN and did not affect ZeroR and 1R at all. Future work could be done in researching why CFS affects certain algorithms more than it does so others.

Reflection

I am greatly attracted by the power of data and the abilities of these predicting algorithm. Each of them is just a normal number, but when they are combine with other data they become more than just a number, and those algorithm, is like tools which could help us to dig the truth from the dataset. Meanwhile from this assignment, i have learned that not every single data is important to the result. We are not only looking for the prediction of diabetes according to the 8 attributes, but we are also trying to find out what attributes are a better predictor of diabetes, i think this is the most important thing i learned from this assignment.