Report
Classification. K-Nearest Neighbours and
Naive Bayes Algorithms.
COMP3308

Introduction

Aim of Study**:**
The aim of the study is to compare Machine Learning algorithms and to see which ones are more accurate and more efficient at classifying a Diabetes data set.
Importance of the problem**:**
The importance of this report can be demonstrated through two following points:

1. Application of algorithms. Finding out the most efficient algorithms and implementing them can yield better results for industrial, academic, and other use cases. The more efficient algorithms are likely to produce less errors and to exploit less time-space resources.
2. Understanding the reasons behind why some classification algorithms are better than others will provide insightful knowledge about which algorithm to use given a specific data and for a specific goal. Although some algorithms might have not performed well in our experiment, given different conditions and requirements they could be beneficial.

Description of DataSet:
The dataset used for classification is called “Pima Indians Diabetes Database”. It shows whether the patient has diabetes given 8 health indicators. Test results for diabetes are represented as a binary “yes”/”no” variable.
In order to avoid overfitting, the curse of dimensionality, and to simplify the data model, we have used a single threaded Correlation-based feature selection (CFS). This feature selection method selects a subset of attributes so that each attribute was highly correlated with the class variable, yet was uncorrelated with other attributes. Best first search strategy with a stopping criterion of five fully expanded non-improving subsets was used to search the space.
Running CFS in Weka produced a reduced list of five attributes, namely:

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

Results and discussion

	ZeroR	1R	1NN	3NN	NB	DT	MLP	SVN
No feature selection	65.1042%	70.8333%	67.8385%	72.6563%	75.1302%	71.875%	75.3906%	76.2021%

CFS	65.1042 %	70.8333 %	69.0104 %	73.3073 %	76.3021 %	73.3073 %	75.7813 %	76.6927 %

Efficiency of Algorithms: All algorithms except for Multilayered Perceptron were completed in Weka or on our own 10 fold in a respectable amount of time.

	My1NN	My3NN	MyNB
No feature selection	68.4928%	71.4472%	76.0034%
CFS	69.75%	73.1771%	76.5702%

ZeroR**:** This is our baseline. The algorithm is simply taking the majority (which is no) and predicts every single outcome to be “no”. If any algorithm does worse than this it is not doing its job of classifying correctly.
ZeroR with CFS: ZeroR has the same accuracy rate in case of CFS feature selection since it CFS feature selection does not change the way the algorithm works.

1R: 1R selects one feature that generates the smallest error rate. In our case 1R has selected Plasma Glucose Concentration as a predictor. This makes sense because many people with Diabetes have very high glucose, which is a key indicator of diabetes. Therefore it makes sense that 70% of the class variables were correctly predicted. Because we don’t just label them all as “no”, we actually use a correlating variable to predict it.
1R with CFS: Since the feature selected as a predictor has also been selected as one of 5 features in CFS, using CFS data set does not change the results of the algorithm.

1NN and 3NN: It was a bit surprising that 1NN did not do as well as 1R. You would think that a patient with very similar backgrounds would have the same condition. However, it is clear that this is not the case possibly because of high variance of outcomes for the closest neighbour. More than 32% of cases are incorrectly classified if you assume that a patient with very similar conditions as a diabetes patient also has diabetes.Taking average of a bigger number of nearest neighbours, for example 3NN, gives a better result by reducing the noise on the classification and creating smoother decision boundaries.
1NN and 3NN with CFS:We can see that both 1NN and 3NN perform better with CFS feature selection since their performance is highly dependent on elimination of irrelevant features.

Naive Bayes: This algorithm did better than we expected it to. Most of the variables contributed somehow to if the patient had diabetes or not, and this was the key reason of why Naive Bayes worked well. Naive Bayes assumes that all of the variables contribute independently to the class variable. Fortunately, most features in this dataset were independent therefore Naive Bayes was able to work efficiently.

Decision Tree: Decision tree took in all of the factors but did not do as well as Naive Bayes. The reason for this is the following: it classified the data based on if each variable was within a certain range. The largest factor is the Plasma glucose concentration. After determining what range the Plasma glucose was within, it took into consideration the BMI numeric. After figuring the range the BMI was in it either classified or took into consideration more variables. In essence, Decision Tree is reliant on Plasma Glucose concentration. All the variables are dependent on a certain range of Plasma Glucose. It is totally unlike Naive Bayes where the variables are independent. It therefore will make more mistakes.

Multi-Layer Perceptron: Multi-Layered perceptron did better than Decision Tree which was expected, and overall MLP did very well. While Decision Tree only looked at at the basic attributes that were fed into the algorithm, Neural Network had several intermediary steps combining the attributes into higher level concepts, with Input layers, hidden layers, and output layers; each with weights. The one downside to the increase in number of steps compared to decision tree was the time it took to run the algorithm. While Decision Tree took a fraction of a second in Weka, MLP took almost 5 seconds to run on both datasets.

Support Vector Machine: Support Vector Machine performed the best out of all of the algorithms. This means that we must have had a clear margin of separation in most of our data, or a large amount of dimensions. It also means that our dataset is relatively not too big because if it was, the training time would have been much longer.

My1NN, My3NN, and MyNB: All classification algorithms written by us performed better than Weka's algorithms without feature selection. In our Weka evaluation we used the default settings which aren't set to yield the best results.
We can see that the batch size of 100 for all three algorithms definitely affects the accuracy rate. Instead of taking the whole dataset the algorithm only considers a 100 examples thus lacks robustness of classification. Additionally, the Weka algorithms only consider 2 decimal places which leads to significant data loss and imprecision of training examples. For instance, 1-Nearest Neighbor in Weka rounds up attribute values to 2 decimal places. As a result, Weka's KNN classifier might have chosen not the best optimal neighbour, since it could be equally distanced away from the testing entry when the attribute values are rounded up.

Impact of CFS Feature Selection:
CFS selected five features out of eight, as mentioned previously. Excluded features are:

1. Number of times pregnant
2. Triceps skin fold thickness
3. Diastolic blood pressure (mm Hg)

Whereas, we can not estimate exactly how much weaker are the above three features in predicting the class variable than the selected five, it is possible to explain the exclusion of those features. For example, the attribute values for number of times pregnant are unlikely to be above 7, and densely cluster around 0-2, in other words, that the vast majority of female patients fall under range of 0 to 2. This attribute in our eyes is not helpful when predicting diabetes.

Triceps skin fold thickness usually can be somewhat related to BMI, therefore can maybe sometimes help predict diabetes, and High Blood Pressure is usually an indicator only if a person has Type 2 diabetes(we have no indication is Pima Diabetes is type one or two). However CFS found that there are variables that are much more useful in predicting diabetes so we do not use these variables.

CFS feature selection improved the accuracy for Weka classifiers. It did not have unnecessary data. The excluded features have a much smaller correlation to diabetes than the included features. Including the number of times pregnant, the skin fold thickness, and blood pressure actually throws off the accuracy leaving us wondering if they actually have any correlation at all to diabetes.

Conclusion

Summarise your main findings and, if possible, suggest future work.

We have found that Support Vector machine is the most effective classification algorithm for this dataset. We have learned that Decision Tree classifies on the assumption that some variables are dependent on one another, and while this is effective for some datasets, it was not effective for this dataset. Naive Bayes proves to be more effective assuming that these variables are independent of one each other and contribute to the Class Variable equally. We learned that Multi-Layer Perceptron takes a larger amount of time than Decision Tree, but it provides more accurate results by providing different hidden layers and weights. We have learned that Nearest Neighbor is not the ideal algorithm for this dataset, and that it does about as well as Decision Tree. The Nearest Neighbor algorithm obviously improves when we increase the sample size and use 3 Nearest Neighbors instead of 1.

We can also conclude that CFS for this dataset improves the results, proving that there are some variables in this dataset that are actually unnecessary when classifying and predicting if a patient has diabetes.

Further work: One suggestion we have is that we expand and actually implement some of the more complicated algorithms like Support Vector Machine and Multi Layered Perceptrons. Another idea that we are almost surprised we didn’t do in this Assignment was trying these algorithms on different datasets. We only tested one dataset. There are plenty of datasets out there where the variables are more dependent on one another (where decision tree would work better). There are lots of datasets that are larger than this one and lots of datasets that are smaller. The point is, different classification algorithms work better with different datasets and in order to see which ones work better for certain attributes, future work is needed.

Reflection

The most important thing we learned in this assignment was probably the implementation of Nearest Neighbor and Naive Bayes. While it is nice to learn about all of these classification algorithms in lectures, we find it more useful to actually implement them so we can learn them inside and out.