COMP3308 Introduction to Artificial Intelligence

Assignment 2

Il Tae Park 440426684
Joshua Vernon 440613608

Contents

1 Introduction

1.1 Aim

The aim of this assignment was to implement the **K-Nearest Neighbour** and **Nave Bayes** algorithm, evaluating them on a real data-set using the stratified cross validation method. Then the performance of other classifiers and feature selection were to be investigated and evaluated using the program Weka. This would allow the strengths and weaknesses of certain classifiers and the programs that implement them to be revealed so as to gain a better understanding of when or when not to use them.

1.2 Importance

The task of comparing different classifiers with specific analysis of **K-Nearest Neighbours** and **Nave Bayes** holds incredible importance in the future of decision based computing. Artificial Intelligence currently solely relies upon efficient and accurate decision making and as decisions can have very high complexity costs the analysis of classifiers is important in knowing when and where to use different decision making techniques. A complete understanding of when classifiers are successful or unsuccessful allows the programmer to bequeath an A.I. with decision making skills akin to that of humans with our ability to use different areas of reason when approaching different tasks.

2 Data

2.1 Description of Data

The data used for the purposes of analysis of the classifiers and feature selection in this assignment is a collection of information from the Pima Indians Diabetes Database. The data was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases and was donated on the 9th of May 1990 by:

Vincent Sigillito
vgs@aplcen.apl.jhu.edu
(301) 953-6231

Research Center, RMI Group Leader
Applied Physics Laboratory
John Hopkins University
Johns Hopkins Road
Laurel, MD 20707

The data was subsequently modified for COMP3308, in March 2015. Missing values were replaced with averages and classes changed to nominal values.
Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The data includes 768 instances with 8 attributes and 1 class each. The class interpretation is that yes means the patient ”tested positive for diabetes”. The class distribution is 268 patients tested positive for diabetes and 500 tested negative.

Attributes (numeric-valued except class):

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 (muU/ml)
6. Body mass index ($weightinkg/(heightinm)^2$)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (”yes” or ”no”)

2.2 CFS Method

CFS (Correlation based Feature Selection) is a measure of evaluating subsets of features based upon the hypothesis that good feature subsets contain features that are highly correlated with the classification, however maintain uncorrelated with each other. CFS subsequently allows the subsets to be selected to be used for analysing the class as a whole.

3 Results and discussion

3.1 Accuracy Results

Table 1: Weka Accuracy Results

Classifier	**ZeroR**	**1R**	**1NN**	**5NN**	**NB**	**DT**	**MLP**	**SVM**
No feature selection (%)	65.1	71.5	67.8	74.5	75.1	71.9	75.4	76.3
CFS (%)	65.1	71.5	69.0	74.5	76.3	73.3	75.8	76.7

Table 2: Implementation Accuracy Results

Classifier	**My1NN**	**My5NN**	**MyNB**
No feature selection (%)	68.5	75.6	75.3
CFS (%)	68.2	75.0	75.9

3.2 Classifier Performance

Using the results collected in Table 1 the performance of the classifiers using Weka’s implementations can be compared for success in predicting the class of the Pima Indians Diabetes data-set.

ZeroR is the simplest classifier which simply chooses the majority class. In this case with 500 instances of ”no” out of 768 instances it has a 65.1% chance of being correct which is a good place to set a benchmark of which all other classifiers should improve upon.

The next classifier analysed is **1R** of which is a simple, yet surprisingly accurate, classification algorithm. It creates one rules that is used for each predictor in the data. it then selects the rule that has the total smallest error and uses that as it's one rule. As such it is likely to have a higher chance at successfully predicting the correct class than **ZeroR** and in this case it was more successful. **1R** had a success rate of 71.5% of which was 6.4% more likely to make a correct decision than **ZeroR**. this suggests that at least some of the attributes are suggestive of the class and the data is therefore useful.

The next classifier analysed was **KNN (K-Nearest Neighbours)** of which as its name suggests chooses the **K-Nearest Neighbours** and classifies new cases based on a similarity measure. In this case **KNN** when $K = 1$ obtained an accuracy of 67.8% of which was not only lower than **1R** but was only 2.7% more accurate than **ZeroR** which just chose the most popular class. This doesn't suggest **KNN** with $K = 1$ will be a very accurate classifier. However, with $K = 5$ the **KNN** classifier made a lot more successful predictions of class. In this case it achieved a success rate of 74.5% of which is 9.4% more accurate than **ZeroR**. This suggests that there may be trends in the data of which can be taken advantage of to maximise successful predictions.

Following the different implementations of **KNN** the **Nave Bayes** classifier was used had the most successful predictions so far with 75.1%. This may be as **Nave Bayes** assumes that attributes are conditionally independent to each other and are normally distributed which gives **Nave Bayes** an advantage over **KNN** for use on this particular data-set.

Finally, **DT (Decision Tree)**, **MLP (Multilayer Perceptron)** and **SVM (Support Vector Machine)** had varying results of 71.9%, 75.4% and 76.3% respectively with **SVM** prevailing as the most accurate classifier at predicting class for the data-set. **SVM** has the major disadvantage of being painfully inefficient to train and as such would not necessarily be the best classifier to use if the data-set was to grow exponentially, however, with this small data-set it is quite appropriate.

3.3 Implementation and Weka Comparison

Weka being a collection of machine learning algorithms for data mining tasks and an industry standard along with other successful programs such as **R** and **TensorFlow** is good calibre of which to pit oneself against when designing decision making algorithms. As can be seen in the accuracy results at 3.1 our implementations of **KNN** and **Nave Bayes** are on par with that of Weka's results. Specifically without the use of **CFS** our implementations slightly improve upon the results achieved by Weka's algorithms. **KNN** with $K = 1$ is improved by 0.7%, **KNN** with $K = 5$ is improved by 1.1% and **Nave Bayes** is improved by 0.2%. Whilst these are small improvements again it has to be stated that in the case of medical analysis small changes such as these could mean the difference of numerous life or death situations. Where our implementation fell down was in the addition of **CFS** of which mildly reduced the success of our predictions for both **KNN** implementations. This may be as the **KNN** algorithm gets its result by obtaining a set of elements closest to the testing element. Whereas the fold is a subset that's divided based only on the classification and as such does not care about distance and therefore does not have its elements distributed equally within a range. As such with the additional use of **CFS** the Weka implementations for **KNN** with $K = 1$ and **Nave Bayes** were more successful than our implementation, though again by minor differences. This does tend to suggest **CFS** used to help determine class for this data-set appears to be unreliable and should be considered more before being included in a final product.

3.4 Feature Selection

It is worth noting that some of the classifiers did have varying success when using or not using **CFS** as feature selection. There was no difference in the performance of **ZeroR**, **1R** and **KNN** with $K = 5$.

However, in the case of **KNN** with $K = 1$ feature selection improved the predictions by 1.2% which is not a large improvement, however, is still the difference of correctly predicting a class for 9 patients and especially in medical prediction's this can make all the difference.

Another significant observation to note is that whenever there was a difference in the success of the predictions using **CFS** it was positive so according to this data-set there is no negligible negative impact of using **CFS**. Specifically in the case of **DT** using **CFS** improved the success rate by 1.4% of which was the highest improvement.

As the feature selection algorithm being used was**CFS** of which is based upon the hypothesis that good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other, it can be suggested that in the cases where the success rate increased that subsets were able to be selected that were in fact highly correlated or at least more correlated to the class than originally.

Finally, it can be seen that **CFS** was detrimental to the success within our implementations and as such further investigation should be done into the affect of using **CFS** to assist with predictions with the classifiers on this chosen data-set.

3.5 General Feedback on Results

The results obtaining from the different classifiers display a range of success from the standard set by the trivial classification used by **ZeroR** of 65.1% to the highly commendable result achieved by **SVM** using **CFS** of 76.7% successful predictions. The range of success is more than 10% and in an area of study such as medicine such a range of difference can be the difference between life or death of millions of patients around the globe. So if an Artificial Intelligence based decision making product was to be produced to have a place in the determination of a patients diagnosis it is very important that studies into which decision making algorithm, classifier and feature selection should be used so as to lower the boundary of incorrect predictions. This analysis however, did not take into consideration the performance of such classifiers and their implementations on much larger data-sets of which would likely be used should more accurate results be desired. As such if larger data-sets are used the results of this analysis could be rendered obsolete or merely assistive in the design of a useful product.

4 Conclusion

4.1 Findings Summary

In conclusion, it can be seen that some classifiers hold an advantage of others especially in consideration based upon this particular data-set. The success of different classifiers had a range over over 10% and as such suggests that the choice of the correct classifier for the data-set is very important and if possible it may be worthwhile to not limit oneself to only one classifier should the computing power and resources be available to you.

It can be tempting to speak metaphorically however since in this case the data is real we can suggest concrete facts about which the success of different classifiers would have upon the patients of which this data speaks of. Given that one or some of these patients had not yet been diagnosed and were awaiting a verdict it is likely that with the assistance of a either our implementation or that of Weka's and the data a person without any medical training would have an average of a 70% chance of predicting a patients diagnosis of positive or negative of diabetes.

This is incredible as currently the resources and man hours that go into training an individual so as to be able to provide patients with such a diagnosis is obscene and as such a software that could provide similar results could inevitably replace some tasks that doctors need to complete so as to perform their jobs.

4.2 Future Work Suggestions

In the future it should be recommended that more work be done on the cross validation of these classifiers and that **CFS** be investigated further so as to ensure it is only used in times of which it will surely have a benefit on the success rate.

Our implementations should also be tested against other data-sets and compared to Weka to see how they hold up as they have a 1% overall a increase and could be be more efficient or the success could be relative to this data-set only.

Another area of which future work should defiantly be performed is in that of allowing artificial intelligence to improve artificial intelligence. In this case it would be to use machine learning to repeatably test the different classifiers with different data-sets and allow the algorithm to learn what classifiers, implementations and feature selections it should use and when to use them.

Finally there should be more research into the minor differences of success rates as when a classifier only has a difference of a few percent it should be known as to why that classifier was more successful and if it would defiantly be more successful for all future cases.

5 Reflection

The most important takeaway from this analysis both researchers agreed upon was that there is incredible opportunity in artificial intelligent decision making algorithms. However, small details in their implementations, the choice of classifiers, feature selection and attribute collection have a huge impact of the success of the decisions that the algorithms make. Therefore there is a lot of work to be done in cataloguing the appropriate use-cases for different classifiers, implementations and modifications to algorithms so as to ensure that in the future when artificial intelligence is used in our every day lives to determine important life changing decisions such as whether or not we have a life threatening disease, should turn left at a set of lights or sell shares on the stock market that it is using the most appropriate decision making tools available to it to do so.

Another takeaway however, is that there is more work to do and that it is most likely the case that artificial intelligence should be used in conjunction rather than a replacement for human intelligence especially in areas of importance such as medicine. This is as until we can collect sufficient data to provide to artificial intelligence humans have years head start on collecting data intuitively and in a way that machines currently don't have the capability of mocking.