COMP3308 Assignment 2

Craig Tiu 9th May 2019

AIM

This study aims to showcase the potential of statistical-based learning on real world applications. In particular, the predictive power of algorithms like Naïve Bayes and K-Nearest Neighbor are highlighted by creating classifiers for the Pima Indian Diabetes dataset. The study could be the key to uncovering crucial information about diabetes not only in the Pima heritage, but perhaps globally as well, improving predictability and allowing for early countermeasures to be taken to prevent it.

DATA

The given dataset is the Pima India Diabetes dataset – each entry being a patient's record. As mentioned in the provided names file, the records are taken only from females of at least 21 years old of Pima Indian heritage. There are a total of 768 records in the dataset, each with 8 attributes and a class, namely:

1. Number of times pregnant
2. Plasma glucose concentration 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 *("yes" or "no")*

Correlation-based feature selection *(CFS)* is also applied to the dataset, which is defined by a heuristic that selects the best subset of the original attributes that are good at predicting the class, and how much they correlate with other features. High class correlation and low correlation with other attributes are signs of a good feature.

The subset of features selected by CFS include plasma glucose concentration, serum insulin, body mass index, diabetes pedigree function, and age.

RESULTS AND DISCUSSION

Following are the accuracy results from performing 10-fold cross validation on each of the algorithms. The values are rounded off to 2 decimal places.

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM	RF
No feature selection	*65.19%*	*70.80%*	*69.75%*	*75.62%*	*74.71%*	*74.58%*	*75.10%*	*76.40%*	*77.44%*
Correlation-based	*65.10%*	*70.83%*	*69.01%*	*74.48%*	*76.30%*	*73.31%*	*75.78%*	*76.69%*	*75.91%*

Results from using Weka 3.8.3 on pima.csv and pima-CFS.csv.

	My1NN	My5NN	MyNB
No feature selection	*68.35%*	*75.39%*	*75.26%*
Correlation-based	*68.23%*	*75.13%*	*76.04%*

Results from my own implementations of the algorithms and validation. (Appendix, Image 1.1)

Apart from ZeroR, 1R and 1-Nearest Neighbor, the classifiers were relatively accurate at around 75% correctly classified examples. CFS had the largest positive impact on the Naïve Bayes algorithms with a 1-2% increase in accuracy, but the largest negative impact for Random Forest with a 1-2% decrease in accuracy. This is also true for my own implementations of some of the algorithms. The accuracies of my own implementations are quite like that of the classifiers from Weka, if not better in some cases.

The general trend from the table seems to be increasing accuracy from left to right. The algorithms on the right are also more complex and take more time to build classifiers and classify new examples. From using the algorithms on the dataset on Weka, ZeroR takes 0.01 seconds to classify while Random Forest took over 3 seconds. This highlights the tradeoff between accuracy and time complexity. In general, more complex algorithms such as the multilayer perceptron will yield better accuracy results but take longer to classify. This tradeoff must be noted as it is highly dependent on the context and situation.

As mentioned previously, CFS did indeed take a subset of the features, bringing it down to a total of 5 attributes, namely the plasma glucose concentration, serum insulin, body mass index, diabetes pedigree function, and age. This selected subset made intuitive sense in that these are the attributes I expected had the greatest impact on the development of diabetes.

Glucose concentration and serum insulin test directly deal with blood glucose levels, which can be assumed to be highly correlated to diabetes. Body mass index is perhaps also be a good indication as it is a result of dietary habits. Diabetes pedigree function provides data about the history of relatives with the disease. Finally, the older someone is, the more they are at risk of/more prone to diabetes.

It can be noticed from the evaluation via 10-fold cross validation that CFS had a minute impact on these classifiers for this dataset. Feature selection had impacted the accuracy results for no more than 2% for both Weka and my own implementations. This is initially confusing, as one would expect the accuracy to improve dramatically given that the *"best"* subset of features is selected by CFS.

While this feature selection does select the best features defined by a heuristic, it may not always greatly improve accuracy. It remains beneficial despite this fact, as now fewer attributes are needed to determine class while maintaining accuracy. This is advantageous for two reasons.

Firstly, for datasets that store thousands or potentially even millions of entries, space is an issue. By using CFS to select only a subset of the most important attributes, the space requirements are greatly decreased, improving data storage efficiency. Secondly, testing for these attributes is made less tedious for the simple fact that there are now fewer attributes to test for. For this example, testing/recording features such as triceps skin fold thickness can be scrapped without worrying that it would impact the accuracy of the results. Further, the reduced number of attributes also impacts the time complexity of the calculations and algorithms overall. Therefore, CFS remains beneficial despite minor changes to accuracy.

CONCLUSION

It can be concluded that at ~75%, the accuracies produced by the algorithms are not quite high enough to be used in the real world. To give some context, 4th-generation HIV tests 28 days after exposure picks up about 95% of infections [1].

The evaluations done previously produced roughly the same number of false positives and false negatives. False positives are quite dangerous as actions taken for diabetes on a non-diabetic person can be disastrous or even lethal. Similarly, false negatives can result in untreated diabetic people, also having serious consequences. An accuracy this low that deals with the health and livelihood of people is therefore not ideal.

Having said that, potential remains for these classifiers. Perhaps performance depends on the given dataset and is limited by the size; allowing training with more examples could increase accuracy. The dataset provided is miniscule relative to existing datasets out there with thousands upon thousands of entries. In addition, there are possibly more attributes that weren't tested, but have high correlation to the given class. Further research can be conducted to identify these features, possibly further increasing accuracy.

Finally, more study can also be done in the predictive nature of these classifiers. Currently, the test measurements are used to identify whether a patient is diabetic or not. The collection of more data and further development of the algorithms could potentially yield a forward-looking test procedure, able to identify the likelihood in developing diabetes in later stages of life.

REFLECTION

The most important thing I've learned from this assignment is that while statistical-based learning can be effective in real world applications, it isn't a one-size-fits-all solution. Simply applying the basic algorithms to implement classifiers does not immediately solve a problem. Further research and development is compulsory if meaningful results are wanted.

Additionally, I am now more inclined to agree that machine learning is powerful and has great potential, which are the reasons why it is extensively studied and developed today.

Appendix

[1] I-base 2016, *'HIV test accuracy, results and further testing'*, viewed 9th May 2019 at http://i-base.info/guides/testing/test-accuracy-results-and-further-testing.

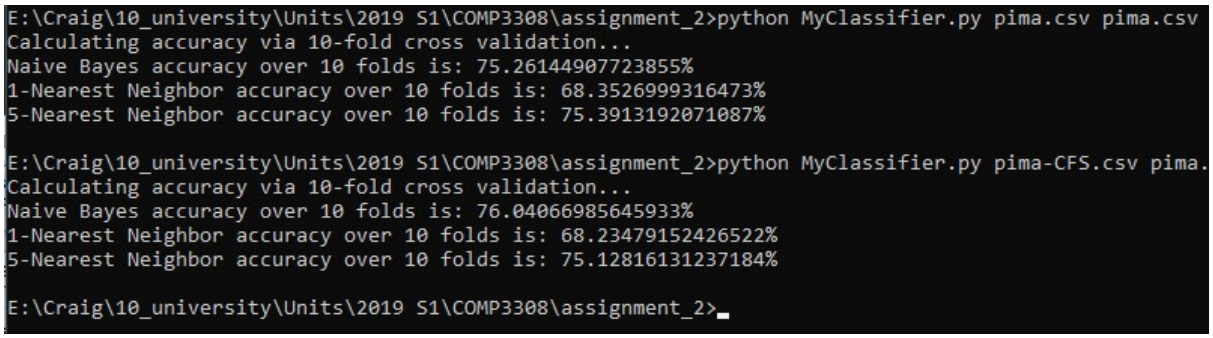

Image 1.1. Accuracy results for Naïve Bayes, 1-Nearest Neighbor and 5-Nearest Neighbor from 10-fold cross validation for pima.csv *(top, no feature selection)* and pima-CFS.csv *(bottom, correlation-based feature selection)*.