COMP3308 – Introduction to Artificial Intelligence

Assignment 2

Name: Sungki Yoo
SID: 440508452

Contents

[1 MARK] INTRODUCTION

- **What is the aim of the study?**

- Diabetes is being treated as a major public health problem nowadays and it is one of the most wide-spread disease. According to the Better Health Channel, around 5.1% of Australians aged 18 years old have diabetes, and this value proportionally increases as the age gets older. So, for this assignment, we are aiming to classify and evaluate the given real data (Pima Indian Diabetes) by implementing the K-nearest neighbor and Naïve Bayes algorithms using the stratified cross validation.

- **Why is this study (or the problem) important?**

- Learning the standard machine learning methods such as the distance based algorithm, K-nearest neighbors and the statistical based Naïve Bayes is important because, through using the application of data mining artificial intelligence (A.I), it is able to quickly and automatically produce models that can analyze growing volumes of real-world complex data with more accurate results.

[1 MARK] DATA – WELL EXPLAINED

- **Dataset – brief description of the dataset**

The data used for this assignment is called Pima Indian Diabetes dataset that is modified by the University for consistency. The original dataset can be sourced from UCI machine Learning Repository, which originally published on 9th of May in 1990, but was modified (replacing missing values with averages, and class were changed to nominal values) in 2015.

There are two types of data files, one for the Pima Indian Diabetes with .data file extension, and one with .name file extension. The data formatted file is the actual dataset for this assignment, and the other one is for the description of the dataset.

For the given dataset, the data contains 768 instances with 8 attributes and 1 class for each of its rows. For the class, there are two classes examined, one with 500 instances of "yes" with positive diabetes and 268 instances of "no" with negative diabetes.

List for each attribute (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 (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”)

The data was nominalized through the utilization of software called Weka to make sure that each attribute of the dataset is in range of 0 and 1 except for the class value, and also used comma as trigger to separate each attribute. This pre-processing was done for clear and better analysis for the dataset.

- **Attribute selection – brief summary of CFS and a list of the selected attributes**

  CFS (Correlation based Feature Selection) is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy. The CFS method is used for this assignment, and to do that, the dataset was transformed from pima.csv data file to pima-CFS.csv data file through using Weka. The Best-First Search algorithm was applied during the processing. From the attributes above, only

  2^{nd}, 5^{th}, 6^{th}, 7^{th} and 8^{th} attributes as well as the class instance from above list were selected for the transformation, and therefore the number of features for testing was reduced.

[6.5 MARKS] RESULTS AND DISCUSSION

Accuracy result from Weka

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM	RF
No feature selection	65.1042%	70.8333%	67.8385%	74.4792%	75.1302%	71.875%	75.3906%	76.8229%	74.8698%
CFS	65.1042%	70.8333	69.0104%	74.4792%	76.3021%	73.3073%	75.7813%	77.0833%	75.9115%

Brief description about the result and the feature selection (CFS):

First of all, for the result using Weka, the accuracies between the dataset, one without the feature selection and one with the CFS algorithm, there was no big intuitive difference. However, the classifiers with CFS always got more accurate results than the one without CFS. This was mainly because, applying the CFS method on each classifiers enabled the classifying algorithms to train the dataset faster by reducing irrelevant and redundant data, hence improved the accuracy of a model.

For example, for the result from the Weka software, except for the **ZeroR**, **OneR** and **5NN** classifiers, in which they resulted 65.1042% and 70.8333% respectively for both non-feature selected and CFS feature selected method, from **KNN** to **Random Forest,** they present a slight difference to each other, depending on whether they are using the feature selection or not.

The rate for the accuracy results from the classifiers that showed the differences:

1NN = 1.1719%
NB = 1.1719%
DT = 1.4323%
MLP = 0.3907%
SVM = 0.2604%
RF = 1.0417%

As you can observer, the least difference was occurred in SVM classifier, whilst the highest difference was found on DT classifier. Meanwhile, 1NN and NB showed the same result.

Accuracy result from my implementation

	My1NN	My5NN	MyNB
No feature selection	69.2549%	73.8295%	74.7454%
CFS	69.4105%	75.3862%	76.2987%

Comparing to the result from the Weka, there seems to have a slight difference in between their accuracy results. Also, the classifiers, NB and KNN that I implemented displayed different accuracy for every time whenever the program was run. Firstly, the difference of the accuracy from Weka and my implementation could be occurred due to the disparity of the implementation for the 10-cross validation. Secondly, getting different result every time I run my program was because of the Collection.shuffle during the cross validation in order to ensure that the test sets do not overlap. For the accuracy from the one I implemented, there was also a difference in their result depending on its contingent upon the CFS algorithm. Here again, the one with the CFS method illustrated higher accuracy than the one without the feature selection.

[2 MARKS] CONCLUSIONS AND FUTURE WORK

From the result, it was effectively evidenced that the utilization of Correlation-based feature selection surprisingly increased the accuracy for each of the classifiers except for some of them. While testing and running the both programs, Weka and my implementation, it was able to observe that, using Naïve Bayes classifier was much faster than the K-nearest neighbor due to KNN's real-time execution. In fact, the running cost for NB takes O(1) in big Oh notation whilst KNN takes O(n^2) where n is the number of the data points, which means that, as the data size increases, plane KNN

classifier usually becomes useless. For overall, the processes above all showed the commendable result and their accuracy was able to be complemented through using the CFS algorithm.

Eventaully, for the future work, in order to boost the accuracy of a model, it is significant to utilize the appropriate feature selection in order to eliminate irrelevant and redundant data form the dataset. Also, adding more data could be another way to improve the accuracy. Through using these methods and information, developing an expert system of diabetes will significantly decrease the healthcare costs via early prediction and diagnosis of diabetes.

[0.5 MARKS] REFLECTION (MEANINGFUL AND RELEVANT PERSONAL REFLECTION)

Throughout this assignment, I was able to learn more about the machine learning method, and about how to approach towards the problems with the knowledge about classification algorithms. I also could be able to learn how to improve the accuracy for the classification, so in future, I could also analysis other problems with better accuracy, hence get an enhanced and improved result.

BIBLIOGRAPHY

- Department of Health & Human Services. (2013, July 31). Diabetes. Retrieved from https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes

- Singha, S. K., & Hassan, S. I. (n.d.). Enhancing the Classification Accuracy of Noisy Dataset By Fusing Correlation Based Feature Selection with K-Nearest Neighbour. Retrieved from http://www.computerscijournal.org/vol10no2/enhancing-the-classification-accuracy-of-noisy-dataset-by-fusing-correlation-based-feature-selection-with-k-nearest-neighbour/