Ross Bollinger

Professor Nathalie Japkowicz

13 October 2018

CSC 480

Assignment 1

**Question 1: What techniques are used in WEKA to deal with the continuous versus discrete attribute issue in the case of C4.5 (J48) and MLP?**

There is a discretization filter in WEKA that can be used to convert continuous data to discrete. When choosing a filter in the WEKA explorer, under the “attribute” folder, there is an option called “Discretize”. To configure this filter, one can simply choose the indices of the attributes to discretize, with the default being to discretize all attributes. To check that this was done correctly, you can click on each attribute and view the details in the “Selected attribute” window. This filter performs binning on the data, with full customizability of the bin range, and the number of bins.

**Question 2:**

Introduction

When discussing the efficiency and performance of unsupervised learning, two attribute types always come up. These two types are continuous and discrete. The two common classifiers used with these two types are C4.5 (or J48), and Multilayered Perceptrons (or MLPs) respectively. However, the question of whether or not these classifiers are actually the most reliable for their respective data types often comes up. In this paper, I plan to test the following related hypothesis: it is best to use classifiers well-suited to the natural knowledge structure of a domain than to convert the domain’s knowledge structure to a less natural one, but one that is appropriate for the classifier in use. I plan to do this by running ten data sets (five with discrete data and five with continuous data) first through their naturally suited classifiers, then convert the data and run them through their less natural classifiers. This will all be done using WEKA (Waikato Environment for Knowledge Analysis).

Data Conversion

The only data needing conversion is the continuous, considering J48 can only be run on discrete data. To do this, I will be using the “discretizer” filter in WEKA on the data sets using the filter’s default value. This filter uses binning to discretize data, with a customizable number of bins. However, I will only be using the default values with this filter when discretizing my continuous data. The default values of this filter create ten bins for each data set.

Experimentation

In this experiment, I will select ten data sets. Five of them will contain discrete data, while the other five will contain continuous data. The data sets are as follows:

Discrete: BalanceScale, CarEvaluation, CongressionalVoting, Nursery, and TicTacToeEndgame Continuous: BreastCancerWisconsin, Ecoli, GlassIdentification, ImageSegmentation, Yeast

First, I will run each of them through their respective natural classifiers and log the results. Then I will convert the continuous data to discrete, then run the sets through their unnatural classifiers and log the results. I will also be recording the kappa statistic results during this experiment.

Results

To evaluate the performance of each classifier, I logged the percentage of correctly classified instances for each data set. This is the most effective method of evaluation because it is the direct measure of how well the classifiers were able to do their jobs. The graphs below show these logs.

**Fig. 1**

Discrete:

Continuous:

I also recorded the kappa statistic with each classifier. A kappa statistic is meant to measure the agreement of prediction between the classification done by the classifiers and the original classification. It is useful for this type of evaluation for similar reasons to the percent correctness values. However, this measure is also factors chance out of the equation. This means that it avoids any data that agrees with a class by chance, rather than perfectly matching with said class. If the kappa statistic is 0, then all agreement was by chance alone, and if it is 1, then all agreement was through perfect matching.

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| κ stats | | | | | |
| Discrete | | | Continuous | | |
| Data Sets | J48 | MLP | J48 | MLP | Data Sets |
| BalanceScale | 0.5773 | 0.8392 | 0.8769 | 0.9086 | BreastCancerWisconsin |
| CarEvaluation | 0.5356 | 0.5911 | 0.5898 | 0.7201 | Ecoli |
| CongressionalVoting | 0.7194 | 0.8893 | 0.8915 | 0.9492 | GlassIdentification |
| Nursery | 0.9227 | 0.9146 | 0.7444 | 0.8611 | ImageSegmentation |
| TicTacToeEndgame | 0.6669 | 0.9264 | 0 | 0.4673 | Yeast |

All of these results only seem to confirm half of my hypothesis. They both show that MLPs are still best to use on continuous data, however they also show that MLPs are best to use on discrete data. These results fail to actually show the benefits of J48 on either type of data.

Conclusion

In conclusion, the hypothesis was only proven partially correct. The results from experimentation show that it was best to use MLPs on their natural data type (continuous). However, the results also show that MLPs were best to use on discrete data. The results do not prove the usefulness of J48 search trees as a classifier at all for either type of data set. This most likely happened due to error during experimentation. Specifically, with the conversion, I could have changed the binning values to be higher, making the data somewhat more specific and less easy for the classifier to process. This would probably have given more precise results, and possibly a different outcome. I also could have looked into more methods of converting the continuous data. This probably would have been a better way to get a better look at the significance of my current results. More methods would have given me results that let me compare which conversion method would actually make a difference.

There are multiple further steps that could be taken with this research. Firstly, going back and testing other methods of conversion as a baseline for an entirely new experiment would be an effective way of finding the right method for repeating this project. Second, while doing this experiment, it was evident that MLPs were taking much longer to run on data sets overall than J48. This is also something to look into as it could show an effectiveness versus efficiency argument between the two classifiers that is worth testing and adding into this project. Finally, after these two steps are taken, another experiment could be done on data sets without classes. This experiment could factor in all the results and conclusions of the other steps and use them to determine the best conversion method for the data and background knowledge on the kinds of results to expect from the J48 and MLP classifiers.