Machine Learning Project 1: Design Document

Trenton Baker b.trent5@gmail.com

Logan Bonney wakeup2early@gmail.com

Bradley White white.brad17@gmail.com

1. Description of the Problem

The purpose of this project is to provide an introduction to machine learning as well as develop an understanding of various machine learning algorithms by predicting how they will work on various real-world datasets and making a comparison between their performance, based on loss functions. A converter was designed to transform these data sets into the proper ARFF format to be interpreted by WEKA (wek).

2. Software Architecture

To be able to use WEKA with these data sets, it is required to convert them into the ARFF format (arf). This could be done manually, but that would prove quite tedious and error prone. A simple program was written in Python, depicted below as a UML class diagram in Figure 1. The converter class parses an input data file, from (uci), and appends attributes and classifications to each line. The new lines are then written to an output file which is in the correct format for WEKA. Since the program contains a single class, a design pattern has been omitted due to simplicity.

3. Design Decisions

Python 2.7 was chosen for the converter because it provides a fast design to implementation life cycle. Furthermore the converter itself is a simple program, therefore it would be easier to understand in Python. The datasets which were chosen were based off of their lack of missing values and they present a range of attributes and instances for evaluation of the algorithms (uci). The datasets have attributes of size: $|a| = \{6, 7, 8, 11, 30\}$ and instances of size: $|i| = \{194, 1728, 4177, 28056, 1025010\}$. The names of the chosen datasets are: abalone, krkopt, car, flag, and poker-hand. Poker-hand does appear to be an outlier within $|i|$, so caution will need to be used if evaluations are made based on a certain algorithm across multiple datasets.

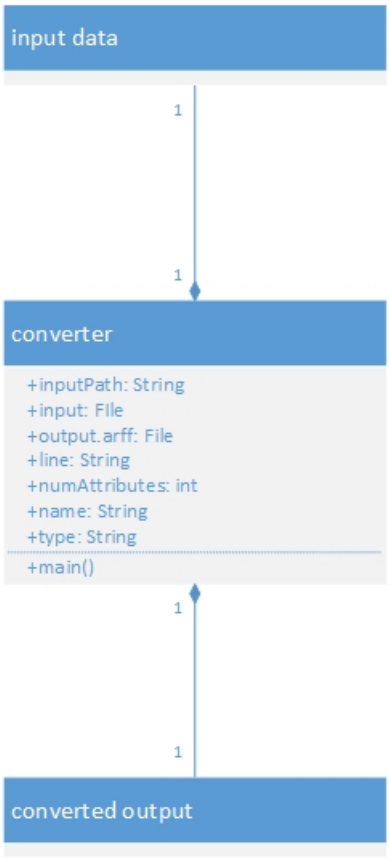

Figure 1: UML Class Diagram for the converter

4. Experimental Design

Each algorithm will be run using cross-validation with ten folds and then the different algorithms will be compared based on the data WEKA generates about time-elapsed, CPU use, false positive percentage, and false negative percentage. The five classification algorithms which will be evaluated are:

1. K-nearest neighbor (IBk) using $k > 2$
2. Naïve Bayes
3. Logistic regression
4. Decision tree with pruning (J48)
5. Support vector machine with a nonlinear kernel (LibSVM)

References

Arff (developer version). URL http://weka.wikispaces.com/ARFF%28developerversion%29.

Uci machine learning repository data sets. URL http://archive.ics.uci.edu/ml/datasets.html.

Weka 3: Data mining software in java. URL http://www.cs.waikato.ac.nz/ml/weka/.