ECS 132 Final Project

Aaron Okano, Anatoly Torchinsky, Justin Maple, Samuel Huang

November 19, 2012

1 Forest Fire

In [Cortez and Morais, 2007], the output 'area' was first transformed with a ln(x+1) function. Then, several Data Mining methods were applied. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using a 10-fold (cross-validation) x 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value: 12.71 +- 0.01 (mean and confidence interval within 95best RMSE was attained by the naive mean predictor. An analysis to the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority. Our goal with this data is to predict the fire size.

2 Parkinson's disease

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.For further information or to pass on comments, please contact Max Little (littlem '@' robots.ox.ac.uk).

Our goal with this data is to try to monitor patients with Parkinson's disease remotely, by simply analyzing their voices on the phone.

3 Beyond ECS 132

Professor Devanbu's paper, "Clones: What is that smell?" assesses the validity of the "stink" that surrounds clones. This reputation stems from the long standing belief that clone's require more project maintenance, as well as their tendency to create code bloat. One of the major problems that people face with software life cycles is maintenance costs, which can require around 80% of the total cost. As a result, Devanbu and others have invested time in research to minimize maintenance costs. One of the simplest ways to lower these costs involves reducing defects in code. Three tests were structured to analyze the relationship between clones and bugs. The first test determines the bug rate of cloned code; the second test compares the bug rate in cloned code to that of regular code; and the last test checks if prolific clone groups are buggier than the non-prolific clone groups. The results and conclusions of each test were formulated using statistical computation and analysis.

The first test was an attempt to discover to what extent cloned code contributed to bugs. The graph plots the cumulative bug convergence on the Y-axis against the clone ratio on the X-axis. According the vertical line, which represents the average clone ratio across all snapshots, both of the projects had that about 80% of bugs have a lower clone ratio than the overall project clone ratio. The test shows that in each case the bugs contained little cloned code.

The second test finds whether clones occur more often in buggy code than elsewhere. A box and whisker plot simply illustrates for all four projects that buggy code had a lower clone ratio. This provides strong evidence that clones are not a large contributor of bugs. To further support this claim, a second table displays the adjusted p values in which a Wilcoxon paired test is used as the null hypothesis and the alternative hypothesis is set to "snapshot clone ratio > bug clone ratio". Since the p values shown are between 0.01 and 0.05, this provides moderate evidence against the null hypothesis in favor of the alternative hypothesis. In each of the four projects, clones were not a major source of bugs. This suggests that clones are less buggy than regular code.

The third test assesses whether prolific clone groups are buggier than non-prolific clone groups. This was accomplished by finding the defect density in prolific clone groups and comparing it to those found in non-prolific clone groups. Figure 2 utilizes the resulting data in a box and whisker plot. The graph shows that in each case the bug density in prolific clone groups is lower than that of the non-prolific clone group. In addition, table three utilizes a Wilcoxon test with the alternative hypothesis set to "defect density in non-prolific groups > defect density of prolific group" provides p values that are below 0.5. These p values reject the null hypothesis in favor of the alternative, providing evidence in that more prolific clone groups are less buggy than non-prolific clone groups.

In each of the three tests, the clones were falsely attributed to a bad trait. Since the four projects used were medium to large open source projects, this data should be applicable in most situations. One area of contention over these results is how cloned code is identified. To address this issue, the tests were applied to two separate data sets. The first data set uses a conservative clone detector, while the second employs a more liberal one. Both detectors require 50 tokens in length to consider a code segment to be cloned, but the liberal detector allows for 1% less similarity than that of the conservative detector. Since analyzing both data sets led to the same conclusion, the difference in detection methods is negligible. Ultimately, the evidence from each test shows that clones may have unfairly garnered a bad reputation.