**Lab**

**Aim**

To learn about imbalanced data and to revise looking up the code in the other labs.

**Before you start**

Load the following packages (download them if needed):

* caret
* RColorBrewer
* scales
* cluster
* rattle
* rgl
* fps
* pvclust
* ggplot2

Download the file synthetic\_imbalance.csv from Moodle. The “class” column will be interpreted as a number. Set it to be nominal like this:

imbData <- read.csv(file="synthetic\_imbalance.csv", header=T, sep=",", row.names=1, stringsAsFactors = T)

imbData$class <- as.factor(imbData$class)

**Exercise 1**

Train and compare 2 different classifiers of your choosing using the same trainControl parameters. I chose C5.0Tree and rpart. Compare their results in enough depth to satisfy yourself that the imbalance makes the dataset tricky to classify.

**Exercise 1b**

Grab the synthetic\_imbalance\_2.csv and load it in the same way. Set a seed of 123 and use 10-fold cross validation to train a classifier. **Do you see any errors or warnings come up?** Given that the dataset imbalance is very high in this dataset, you might expect some of the folds to contain no samples of the minority class. Try again with synthetic\_imbalance\_3.csv – this is an even more imbalanced dataset. Do you see any more warnings?

**Exercise 2**

Find a suitable number of clusters for the dataset using a suitable method. Cluster the dataset using an appropriate method and visualise. Do the clusters exactly match the classes?

**Exercise 3**

Add the clustering information in to the dataset and train a classifier (rpart for nice visualisation, C5.0Rules for nice display) to distinguish between the classes. Does the classifier utilise the class label?