Predicting one-year postoperative death rates following thoracic surgery

Craig Heptinstall Crh13- 110005643
SEM6420
Institute of Computer Science
Aberystywth University

***Abstract*—Predicting medical data is becoming more and more vital towards helping reduce death rates, or reducing risks in medical procedures. With the use of tools such as WEKA [1], the ability to build machine learning classifiers has become easier, and has helped produced more reliable results. This paper looks at death rate prediction following thoracic surgery, and compares different means of prediction and configurations on a given test set of data.**

I. Introduction

Thoracic surgery is a means of lung resection, and is performed in order to remove part or all of a lung from a patient who is suffering or has suffered from lung cancer [2]. The data provided for training a machine learning approach to predicting the death rate one year after surgery contains 300 patients, each of which hold several attributes, including age and Risk1Yr. The most latter of these states is the example patient survived (indicated by a 0) or did not (indicate by a 1). The aim of the classifier trained at the end of this report is to be able to predict the survival rate in conjunction with an identical test set of data.

For easier statistical reading, and the ability to easily try a range of classifiers on the training data, Weka was chosen to perform experiments and training. Weka allows pre-processing of data, classifiers and filters to be applied, and also provides good visualisation of training results, such as ROC curves. Throughout this paper, both graphs generated within Weka, and ones created from comparing different classifiers will help to reinforce result conclusions.

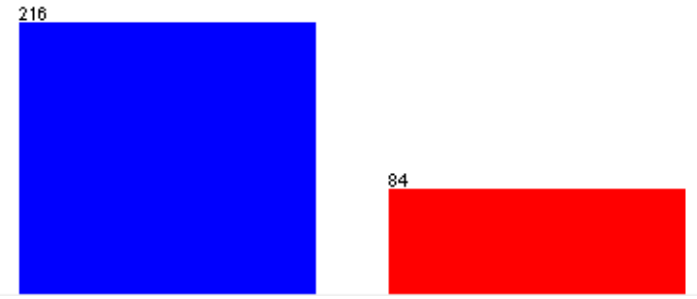
To visualise data quickly during the fine tuning and selection of classifiers during the start of this project, the Weka Explorer application will be used, and then will be moved onto the experimenter during latter parts of the project in order to quickly run multiple experiments at once. Finally, the Weka Knowledge flow application will allow comparisons of the ROC curve of multiple solutions in order to choose the most suitable set up for predicting the test data results.

As stated in the assignment, more than one classifier will be selected and compared to see which one best suits predicting this type of data. Following that, more comparisons will be made looking at changing a range of hyper-parameters for chosen classifiers.

A. Data pre-processing

In Weka, and in any machine learning approach, some data pre-processing should be performed. Because the data set is real-life based, data is not always complete, or follow a linear pattern. Due to this, an issue known as class imbalance can hinder the performance of learning systems, and means that one instance of an attribute (in this example, surviving patients) is the majority sample causing a bias when training the machine learning algorithm. On first inspection of the data in the training set, it can be seen in Figure 1 that the Risk1Yr attribute holds far more surviving patients that ones who did not.

Fig. 1. Visualisation of training set risk year 1 mortality rate. Left: Survivors, Right: Deaths



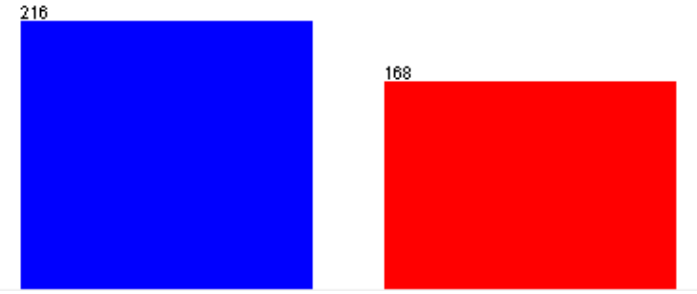
There are two means of settling imbalances in data however [3]:

1) Under-sampling where the size of the samples are reduced proportionally
2) Over-sampling where the size of samples are increased proportionally

In the case of Weka, there are filters available within the pre-processing stage of experiments allowing the addition of under or over-sampling. In this case, the SMOTE (Synthetic Minority Over-sampling Technique) [4] plugin for Weka offered an easy means of re-sampling data based on a given set of parameters. In an accompanying paper [5], SMOTE is described as creating synthetic examples rather than replacements that are generated in feature space rather than data space. In the Weka library, synthetic samples are created by using a range of nearest neighbour minority samples, and placing new samples

between them. By default, the SMOTE filter uses five nearest neighbours to create new samples, and as this is recommended for this size dataset, this will remain the selected value.

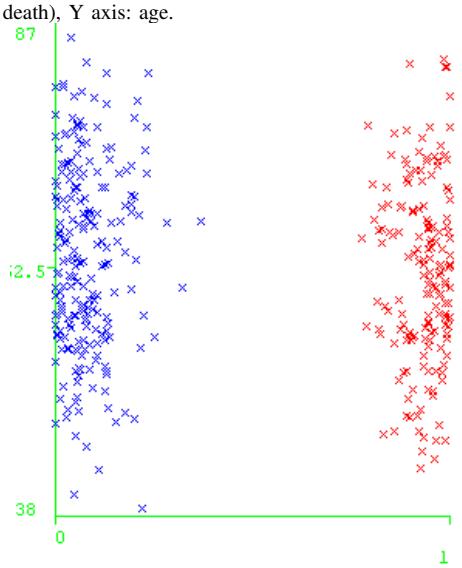
Fig. 2. Risk one year data after SMOTE filter applied.



Once the SMOTE filter was applied, there was a clear view of change in the proportions of the survival and death rates. As Figure 2 shows, over-sampling has been applied to 200%, which has meant the minority of samples showing death rates only has 50 samples less than the majority as opposed to the 138 sample difference previously.

In addition to balancing data, viewing the data in Weka explorer on scatter gave good insight into any data trends, and also gave hints into which attributes carried more weighting towards the risk attribute than others. At first, both age and risk were plotted against each other to see if there were any trends towards being younger meant better survival rates (as shown in Figure 3). In order to view the scatter plot well enough, Jitter was added in Weka to plot points further away from each other.

Fig. 3. Scatter plot comparing age and survival rates. X axis: Death rate (0- survival, 1- death), Y axis: age.



Although the plot looks fairly evenly spread on age for both survival and mortality rates, there is a slight higher number of patients who survived aged around 50 than those who were older.

By looking at the scatter plots for various combinations of data, dimensionality reduction is something that can be used to both increase the accuracy and speed of the resulting algorithm. There are various means of dimensionality reduction, such as PCA (Principle Component Analysis) and SOM (Self Organising Maps) /citedimensionality, whilst in Weka, attributes can be ranked, in order to see which ones carry the most important in a data set.

References

[1] University of Waikato, *Weka 3: Data Mining Software in Java*, Machine Learning Group at the University of Waikato, *http://www.cs.waikato.ac.nz/ml/weka/*, 2013.

[2] Boston Medical Center, *Center for Thoracic Oncology*, Boston University school of medicine, *http://www.bmc.org/thoraciconcology/treatments/lung-resection.htm*, 2014.

[3] M. Rahman, *Machine learning based data pre-processing for the purpose of medical data mining and decision support*, Hull University, 2014.

[4] Microsoft, *SMOTE*, Microsoft Aure Modules, *https://msdn.microsoft.com/en-us/library/azure/dn913076.aspx*, 2016.

[5] N. Chawla et al., *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research 16 (2002) 321357. 2001.

[6] S. Reid, *Dimensionality Reduction Techniques*, Turing Finance, *http://www.turingfinance.com/*, 2014.