Edinburgh Napier UNIVERSITY

Data Analytics- Coursework Report

Emma Parsley
40206111@live.napier.ac.uk
Edinburgh Napier University - Data Analytics (SET09120)

Abstract

This projects aim is to prepare, analyse and visualise the data historically recorded by a bank in Germany. The data will be analysed using algorithms such as classification and regression.

Keywords – data, preperation, mining, cleaning, visualisation

1 Introduction

The data analysed in this project will be the credits data set given to us for this poject. This data set contains 1000 rows each with 13 attributes. The data will need to be cleaned and prepared, before it can be analysed. OpenRefine will be used to clean and prepare the date, the program Weka will be used to analyse the data and it will be visualised using R.

2 Preparation

2.1 Cleaning

In order to clean the data the dataset was opened in OpenRefine. The most obvious problem with the data was that some of the fields under the attributes; checking_status, credit_history, saving_status, personal_status and job, had single quotes around them. This was easily rectified using the replace command in OpenRefine to transform all the cells in these coloums, replaceing the single quote with and empty string.

The purpose attribute contained many values that where the same except they were written or spelled differently. Any values that simply had single quotes around them were fixed in the previous step. A few values needed fixed so they were all spelt correctly and others just needed fixed so their use of case was consistent (e.g Education needed to become education).

From sorting the credit_amount attribute from highest to loweset it becomes clear that there are 8 rows of data with what seem like ridiculously high amounts of money for it's purpose, it seems likely that there were extra zeros at the end, however there is no way of knowing how many of the zeros were accidental so those rows were removed from the data set.

There were also many ages that were negative which was fixed by taking the absolute value of all the ages. There were also a few impossible ages, values less than 1 that were assumed to be correct without the "0." in front, for example 0.24 would become 24. After all this was sorted there were only 2 unlikely ages left 1 and 6, it seems likely that there was a number missed out when these were entered but as there as no way of knowing if these values should be 26 or 60 or something else these sets of data where removed.

The existing_credits attribute had many values that seemed incorrect. For example there shouldn't be able to be a value that is isn't in the set of natural numbers, so for example 1.1 was replaced with 1. It also seemed clear that someone shouldn't be able to a value that isn't a single digit, as anything greater than that seems excessive, so anything greater than 9 was replaced with it's greatest most common digit, for example 101 became 1.

Two entries have the job field filled out as "yes",

this isn't a valid entry so these were replaced with the same job as other similar entries, for example the 37 year old single male that is using the loan for education, their job was replaced with "unskilled resident" as this is the same as someone with a similar entry and it does make sense that someone that wants money for education is likely to be unskilled.

Many entries in the num_dependants attribute where written out rather than entered as numbers, for example 1 was written as "One", these entries merely needed to be fixed to only be written as a number.

2.2 Preparation

To prepare the data to be used it needed to be converted into .arff format so it could be opened in Weka. OpenRefine has the ability to template how it will output, so the data can be exported straight from OpenRefine to a .arff compatible file. All numerical attributes can be removed from within Weka to save a nominal only version of the database.

3 Analysis

4 Visualisation

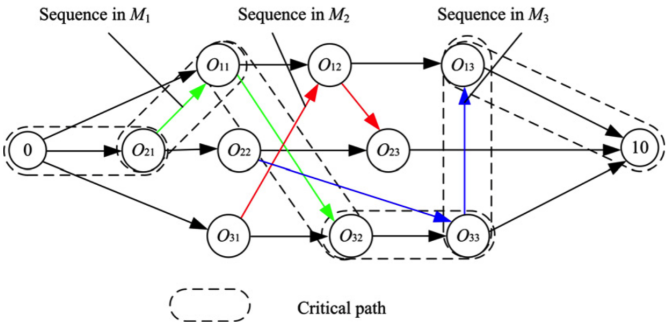

Figure 1: **ImageTitle** - Some Descriptive Text