PREPARING THE "DESERT ISLAND DISCS" ARCHIVE FOR DATA MINING

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2012

By
Elena A. BĂLUŞESCU
School of Computer Science

Abstract

This project seeks to retrieve the archive of the BBC radio 4 show, "Desert Island Discs" and to complete it with metadata describing the guests and their music choices. The main goal is to explore the dataset and investigate the possibility of predicting the music choices of future guests. The study is based on the hypothesis that metadata describing past castaways and their choices on the programme is enough to predict the choices of future guests.

The study is divided in three main parts, metadata collection, data preprocessing and data mining. The project starts by retrieving the online archive of the programme and expanding it with metadata about castaways and music records, collected from DBpedia and MusicBrainz. Several methods of data cleansing were applied throughout this study to ensure that high quality data was used in data exploration. After collection, the dataset was preprocessed and used as input for data mining. The final goal of this project is to investigate the possibility of predicting the musical choices made on the show. To meet this goal, a set of experimental predictions was proposed, namely determining the "castaway's favourite" record and predicting one feature of the favourite record. The model built used the Naive Bayes algorithm in the classification process and return poor results for the rare classes. Performance was improved by adding a cost matrix to the initial model. Classifier's performance was analysed and compared through the precision, recall and F-measure values measured. The results obtained vary from $F = 0.219$ to $F = 0.369$, for the rare classes.

The overall poor results returned in the classification process highlight that the dataset was not appropriate for making accurate predictions. The main challenges identified throughout this research, that might have led to the unsatisfying results, concern information availability, the quality of data, resource identification and data preparation. However, further pathways are proposed for acquiring a more rich and valuable dataset that might offer solutions to the machine learning challenges.