DATA ANALYSIS REPORT

National Research University Higher School of Economics Moscow, Russia

DIMAS MUNOZ MONTESINOS

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The motivation for this work is to apply FCA (aka. Formal Concept Analysis) classification techniques to an UCI dataset [?]. In particular, we will work with the data of a bank marketing campaign. We will measure the accuracy of FCA to predict whether an object should be classified as a positive class or as a negative one.

* *Department of Biology, University of Examples, London, United Kingdom*
[1] *Department of Chemistry, University of Examples, London, United Kingdom*

1 INTRODUCTION

Formal Concept Analysis (*FCA* in short) is a method mainly used for deriving implicit relationships between objects described through a set of attributes on the one hand and these attributes on the other. The data are structured into units which are formal abstractions of concepts of human thought, allowing meaningful comprehensible interpretation. Thus, FCA can be seen as a conceptual clustering technique as it also provides intensional descriptions for the abstract concepts or data units it produces.

In this work, we will use an open dataset from UCI related with direct marketing campaigns of a Portuguese banking institution [**?**]. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed or not.

The classification goal is to predict if the client will subscribe a term deposit. This is encoded as "yes" or "no" in the column "y" of our CSV file.

2 DATA PRE-PROCESSING

2.1 Bank telemarketing data

This data was recolected by Moro et al. [1]. In order to work with the information of the dataset, first we need to pre-process it. In particular, we need to binarize it.

It is important to know that this dataset is **not balanced**: there are 521 positive cases and 4000 negative cases.

2.1.1 *Numerical values*

The *age* will be separated in three intervals: young people (< 27 years old), adult people (from 27 to 65) andold people (above 65).

We have no information about *balance*, so we will separate it into positive and negative values.

For the *campaign* and *previous* columns, we will consider two intervals: $[0, 4]$ and $[5, +\infty)$.

The *pdays* will have two intervals as well: $[-1, 100)$ and $[100, +\infty)$. Note that -1 represents that there wasn't any previous contact.

2.1.2 *Categorical values*

Each value has its own binary column.

2.2 Excluded values

There are some attributes with too many unknown values. This is the case of *contact* or *poutcome*. They won't be considered in our analysis.

We have excluded *job* because of its complexity, although we are going to reduce the precision of our analysis. It would increase the execution time in our analysis.

Finally, we have removed the rows with *unknown* as value of *education*.

3 AGGREGATION FUNCTION

Let G a set of all objects. We get some arbitrary sets G_+ (positive examples) and G_- (negative examples) such that $G_+, G_- \subseteq G$ and $G_+ \cap G_- = \varnothing$.

The objects $G_\tau = G \backslash (G_+ \cup G_-)$ are called undetermined examples, and our goal is to classify them as positive or negative ones.

We need to define our support function ($s : G_\tau \to [0, 1]$):

$$s_\epsilon(g_\tau) = \frac{|g_\tau \cap G_\epsilon|}{|g_\tau \cap G|} \text{ where } \epsilon = \{+, -\}$$

Then, our aggregation function will be as follows:

$$agg(g_\tau) = \frac{\sum_{g_+ \in G_+} |g'_\tau \cap g'_+|}{\max\limits_{g_+ \in G_+} |g'_+| \cdot |G_+||} - \frac{\sum_{g_- \in G_-} |g'_\tau \cap g'_-|}{\max\limits_{g_- \in G_-} |g'_-| \cdot |G_-||}$$

APPENDIX

Attributes in file

The CSV file provides different attributes which allows us to perform the analysis. Here is the list of attributes that we are going to use:

1. *age* (numeric)
2. *job*: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. *marital*: marital status (categorical: 'divorced', 'married', 'single'; note: 'divorced' means divorced or widowed)
4. *education* (categorical: 'primary', 'secondary', 'tertiary', 'unknown')
5. *balance*: no information provided about this attribute.
6. *default*: has credit in default? (binary: 'no', 'yes')
7. *housing*: has housing loan? (binary: 'no', 'yes')
8. *loan*: has personal loan? (binary: 'no', 'yes")
9. *contact*: contact communication type (categorical: 'cellular', 'telephone', 'unknown')
10. *campaign*: number of contacts performed during this campaign and for this client (numeric, includes last contact)
11. *pdays*: number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
12. *previous*: number of contacts performed before this campaign and for this client (numeric)
13. *poutcome*: outcome of the previous marketing campaign (categorical: 'failure', 'other', 'success', 'unknown')
14. *y*: has the client subscribed a term deposit? (binary: 'yes','no')

REFERENCES

[1] P. Cortez S. Moro and P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*. Elsevier, 2014.