DATA ANALYSIS REPORT

National Research University Higher School of Economics
Moscow, Russia

DIMAS MUNOZ MONTESINOS

CONTENTS

ABSTRACT

The motivation for this work is to apply FCA (aka. Formal Concept Analysis) classification techniques to an UCI dataset [2]. In particular, we have been working with the data of a bank marketing campaign. We have measured the accuracy of FCA to predict whether an object should be classified as a positive class or as a negative one.

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 have used an open dataset from UCI related with direct marketing campaigns of a Portuguese banking institution [2]. 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" in the 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 pre-processed it (in particular, we have binarized it).

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

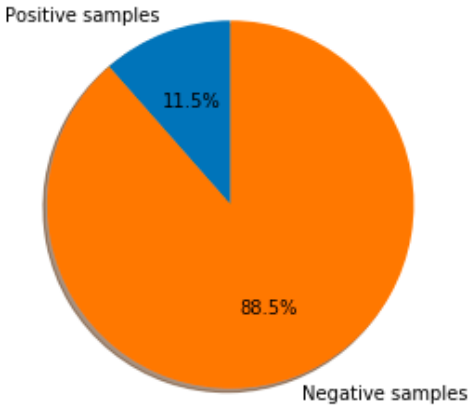

Figure 1: Pie chart of distribution of samples in the dataset.

2.2 Categorical values

Each value has its own binary column. For example, the column *education* has been split in three: *education_primary*, *education_secondary* and *education_tertiary*.

2.3 Numerical values

We have been separated *age* in three intervals: young people (below 27 years old), adult people (from 27 to 65, both values included) and old people (above 65 years old).

In the case of *campaign* and *previous* columns, we have considered two cases: 0 and > 0.

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

2.4 Removed columns and rows

The dataset contains many unknown values in different columns, such as *contact* or *poutcome*, so we are going to remove those columns from the analysis. In a similar way, we have removed the rows with *unknown* as value of *education*.

The reason behind this is that there are only $\sim 4\%$ rows with an unknown value in *education*. In contraposition, there are more than 29% and 80% rows with an unknown value in the columns *contact* or *poutcome* respectively.

Additionally, in order to simplify our dataset, we have excluded *job* because of its complexity, although we are going to reduce the precision of our analysis. Note that it may be easier to analyse this field if we use pattern structures instead of binarization.

We have removed the column *balance* as well. This has been motivated because of the poor results that we have obtained in the analysis.

3 AGGREGATION FUNCTION

Let G a set of all objects. We select some arbitrary sets G_+ (positive examples) and G_- (negative examples) such that $G_+, G_- \subseteq G$ and $G_+ \cap G_- = \varnothing$ [3]. 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.

To classify the objects as positive or negative ones, we use hypotheses. We say that an hypothesis h is positive (idem. negative) if $|\{h : h \subseteq g_+\}| > |\{h : h \subseteq g_-\}|$ for all $g_+ \in G_+$ and $g_- \in G_-$.

Let h^+ and h^- be the hypotheses to classify an item $g \in G_\tau$. The positive hypotheses H^+ (idem. negative hypotheses) are calculated by intersecting all test objects with all objects from G_+.

Additionally, we consider only hypotheses such that the total number of common attributes with respect to some object g should be greater or equal than a fixed value α. More formally, let $g_\tau = \{x_1, ..., x_n\}$ and $g_\epsilon = \{y_1, ..., y_n\}$. Suppose $z_i = 1$ if $x_i = y_i$ (otherwise $z_i = 0$). Then:

$$H^\epsilon = \{h^\epsilon : h^\epsilon = g_\tau \cap g_\epsilon \text{ and } \sum_n z_i \geqslant \alpha\} \text{ where } \epsilon = \{+, -\}$$

We use the hypotheses to compute the support of every object:

$$s_\epsilon(g) = \frac{|h^\epsilon \subseteq g|}{|G_\epsilon|} \text{ where } \epsilon = \{+, -\}$$

We define our score function as follows:

$$score(g) = s_+(g) - s_-(g), \forall g \in G_\tau$$

Then, an undetermined object $g \in G_\tau$ is classified as positive class if $score(g) > 0$. Otherwise, it is classified as a negative one.

4 ANALYSIS

We split the dataset: 2/3 of the data to generate the hypothesis and 1/3 to test them.

	True pos.	True neg.	False pos.	False neg.
Value	85	942	390	90
Ratio	0.4857	0.7072	0.2928	0.5143

Using the aggregation function, we have classified each intent g_τ from the test set as a positive or negative class. Then, we compared with the real value and we got the following results:

Also, we have computed other measures during the analysis:

- Accuracy: 0.6815
- Precision: 0.1789
- Recall: 0.4857
- F1 score: 0.2615

5 CONCLUSIONS

In general, we have obtained poor results when we try to predict *true* classes. Since the ratio is $\sim$ 50%, we cannot be certainly sure whether a customer will suscribe to the promotion or not. In the case of predicting *negative* classes, we got better results, but the ratio is too low as well.

This fact could be caused for different reasons: binarization instead of pattern structures, the chosen intervals are not good, ... Or even the information of this dataset could not be useful at all. It is possible that we cannot predict if a customer will suscribe to the promotion due to other reasons which are not present in the dataset.

In my opinion, we should study which information from the dataset is useful (and which is not) to predict *positive* and *negative* classes. Also, we should study which information should be added to the dataset.

Also, we must take into account that the data is unbalanced, thus the accuracy is not representative at all. We can see that the accuracy value is $\sim$ 68% due to the algorithm has predicted more *negative* classes correctly. To solve this problem, we should find another formula based on weights to measure the accuracy.

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')

Remark: we have not used the *duration* column of the dataset (which is the duration of the last call), although it could improve our analysis. This is because we already know the result after the call finishes.

REFERENCES

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

[2] S. Moro, P. Cortez and P. Rita. *Bank Marketing Data Set*. Published at `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`, Feb. 2012. [Accessed on 28-11-2019]

[3] Dmitry I. Ignatov *Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields.* [arXiv:1703.02819v1]