

CONSULTING THE ORACLE: CAN DATA MINING OVER GEM DATASET TELL SOMETHING ABOUT UNEMPLOYED ENTREPRENEURS?

Francisco Javier Melero Rus, Departamento de Lenguajes y Sistemas Informáticos,
Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional,
DaSCI Universidad de Granada, fjmeler@ugr.es

María del Mar Fuentes Fuentes, Departamento de Organización de Empresas I,
Universidad de Granada, mfuentes@ugr.es

Área: I.1

Keywords: data mining, unemployment, gender, GEM, entrepreneurial intention,

1. GENERAL OVERVIEW AND MOTIVATION

In a recent editorial, Schwab and Zhang (Schwab and Zhang, 2019) highlight the new possibilities that Big Data technologies offer to the entrepreneurship researchers in order to develop new knowledge corpus, although this new paradigm implies to walk nonusual tracks and be open-minded to new data science related techniques.

Inspired by this recommendation, in this paper we present a different approach from the classic research workflow. We do not start from any hypothesis, but from a question that intends to better know the individuals, as a kind of oracle over GEM data: “*What features best describe those unemployed people that answer YES to the questions about future entrepreneurial activity? Furthermore, are there gender differences?*”. In this work we use a data mining technique, the association rules, which will allow us to find the most common relevant answers between the unemployed people that intend to start a business in the following 3 years from the survey or that are at the moment of the survey starting a business, and from that features retrieve the answer to our questions.

2. METHODOLOGY

To answer our questions, we will use the Spain national level APS survey of the year 2015. This survey has 24.300 valid answers, 1306 of them answering YES to the question: “*You are, alone or with others, expecting to start a new business, including any type of self-employment, within the next three years?*” (FUTSUP=1). We want to know the most relevant features of those affirmative respondents that are unemployed (GEMOCCU=6), in order to find their motivations.

Association Rules

Association rules are one of the most used data mining techniques to extract interesting knowledge from big databases, as they identify and represent dependencies amongst elements (items) of a database where we don't know the class of each entry (Agrawal 1993), which is named as unsupervised learning. The knowledge that association rules provide allows to better understand the processes that generated the data, in our case to better understand the unemployed people and their motivation to start an entrepreneurial project.

In an association rule ($X \rightarrow Y$), X is the antecedent and Y the consequent and represents that there is some kind of relationship between X and Y, being X and Y disjoint itemsets.

Some measures of quality are:

- The *support* of an association rule is the frequency of the itemset (X,Y).
- *Confidence* of an association rule is defined as $support(X \rightarrow Y) / support(X)$.

How to apply this technique to find answers to our question? We are going to find the relevant rules that have the largest *confidence* values amongst those whose consequent is {FUTSUP,1}. In other words, we are going to seek the most common answers that are given by those unemployed respondents that manifest their intention to start a business in the following three years.

Data preprocessing

After removing dummy and computed variables, the *apriori* algorithm (Borgelt and Kruse, 2002) generated 437 frequent itemsets that satisfy the quality requirements (support 0.01, confidence 0.2), and then we have created some subsets to work with:

- futsup1: rules whose consequent is only the {futsup=1} itemset;
- male.futsup1: rules that have {gender=1} in the antecedent and {futsup=1} in the consequent;
- female.futsup1: rules that have {gender=2} in the antecedent and {futsup=1} in the consequent.

3. MAIN RESULTS

GEM 2015 data tell us that *“Expect to start a business within 3 years...”*

- ... half of the unemployed people that answered that they were right now doing something to start a job {bstart=1}.
- ... approximately one third of unemployed university graduates {UNEDUC=5} that *“know someone that has already started a business”* {knowent=1}, or that consider that *“have the knowledge, skill and experience required to start a new business”* {suskill=1} and *“do not fear to fail”* {fearfail=0}.
- ... one fourth of unemployed men where {suskill=1}, that *“the fear of failure would prevent them from starting a business”* {fearfail=1}, that consider *“that most people would prefer that everyone had a similar standard of living”* {equalinc=1} and think that *“most people consider starting a new business is a desirable career choice”* {nbgoodc=1}.
- ... one fourth of 45-54 unemployed people {age9c=5} that answered {suskill=1} and {nbgoodc=1}.
- ... 27% of unemployed people that answered {knowent=1}, {suskill=1}, {nbgoodc=1} and consider that *“those successful at starting a new business have a high level of status and respect”* {nbstatus=1}

If we disaggregate the dataset into men and women, trying to find similar or different behaviours, we found that most important features in men are university grade, entrepreneurial abilities and knowledge, and the social entrepreneurial network, apart from detecting opportunities in the environment or lack of fear to failure. On the other hand, women declare mainly fear to failure (but it does not avoid them to intend to become entrepreneurs) and a surprising lack of detection of opportunities, but consider themselves with entrepreneurial skills and a good social perception of entrepreneurs and entrepreneurial career.

4. CONCLUSIONS

We found an anomaly in the FUTSUP variable. In our opinion, the fact that 50% of unemployed people declare to be starting a new business now and also that they are thinking in starting a new one in the future 3 years, may point out a misunderstanding during the data collection, as it seems that an unemployed person is not the ideal profile for a serial entrepreneur, given its income or socio-economic context.

Data tell us that male potential entrepreneurs are more influenced by external references and self-confidence, whereas female potential entrepreneurs plan to start a new business only when there is a low risk of failure, as they state that the risk of failure may prevent them to start the business.

Association rules, as other data mining techniques, may help researchers to find new features of individuals based on their answers to GEM survey. Although well-known behaviours will appear as the rules with greatest confidence (which validates the tool) there might also arise new issues that add knowledge to the field.

REFERENCES

- Schwab, A., & Zhang, Z. (2019). A new methodological frontier in entrepreneurship research: Big data studies.
- R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, Proc. SIGMOD, Washington D.C., 1993, 207-206.
- Christian Borgelt and Rudolf Kruse (2002) Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics (COMPSTAT 2002, Berlin, Germany)* Physica Verlag, Heidelberg, Germany.