Categorizing nonprofit organizations according to their field of activity: A discussion of rule-based categorization and machine learning, and recommendations for implementation

Julia Litofcenko, WU Vienna University of Economics and Business

Dominik Karner, WU Vienna University of Economics and Business

Florentine Maier, WU Vienna University of Economics and Business

In this research note we discuss the two basic computational methods available for categorizing nonprofit organizations (NPOs) according to their field of activity based on textual information about these organizations: (1) rule-based categorization and (2) pattern recognition by using machine learning techniques. These methods provide a solution to the widespread research problem that quantitative data on the activities of NPOs are needed but not readily available from administrative data, and that manual categorization is not feasible for large samples. We explain both methods and report our experience in using them to categorize Austrian nonprofit associations on the basis of the International Classification of Non-Profit Organizations (ICNPO). Since we have found that rule-based categorization works much better for this task than machine learning, we provide detailed recommendations for implementing a rule-based approach. We address scholars with a background in data analytics as well as those without, by providing non-technical explanations as well as open-source sample code that is free to use and adapt.

Introduction

The increasing availability of large amounts of rich and growing administrative or otherwise process-generated data, often referred to as big data, has prompted scholars to consider new ways to use these data for research on nonprofit organizations (NPOs) and civil society (see, for example, Lecy & Thornton, 2016; McDonnell & Rutherford, 2018). One important piece of information concerns NPOs’ fields of activity. Information on NPOs' activities is often of interest in itself, e.g. for mapping purposes, or is needed as a control variable in studies to investigate causal relationships. Unfortunately, however, many available data sets do not contain such information in readily usable form, either because categorization by fields of activity is completely missing, or because it does not have the desired quality (see, for example, Grønbjerg & Paarlberg, 2002:588 on consistency problems with NTEE codes in IRS data in the U.S.). The research task of complementing existing data sets of NPOs with an additional variable that indicates NPOs' main field of activity (or all their fields of activity, for more detailed analyses) is therefore relatively common, but there is still no shared understanding of methods to accomplish this task.

The aim of this research note is to contribute to a common understanding of computational methods for categorizing NPOs according to their field of activity, based on information about the NPOs in text form (e.g. the name of the organization, or written descriptions of the organization's activities). We do so by outlining the two basic computational methods available for this task: Rule-based categorization, and categorization based on machine learning (Zhai & Massung, 2016:300-302). We explain the ideas behind both methods, and report our experience in applying them to categorize the full population of nonprofit associations in Austria based on the International Classification of Nonprofit Organizations (ICNPO, see Salamon & Anheier, 1992). As we find that rule-based categorization performs much better than machine learning in terms of accuracy and transparency, we provide detailed recommendations for efficiently implementing a rule-based approach. We conclude with summarizing our reasons for endorsing a rule-based approach, as well as acknowledging its disadvantages.

It should be noted that we implemented approaches that assign each NPO to one ICNPO category, based on its apparent main activity. However, the approaches could also be adapted to capture several fields of activity. We implemented a categorization into subgroups at the second level of ICNPO. An exception was made for the major activity group of health (group 3). Here we categorized only at the level of the main activity group, because for many organizations the text data did not allow for more precision (e.g. discerning whether a health NPO runs mainly hospitals or mainly nursing homes).

Rule-based categorization

Rule-based categorization is based on the manual creation of IF-THEN rules for categorization. A simple example of such a rule is: IF the organization's name includes the word "fan club", THEN the organization is assigned to the ICNPO-category "1 300 – other recreation and social clubs". As suggested by Zhai and Massung (2016:301), rule-based classification is likely to work well if the following criteria are fulfilled: (1) Categories are clearly defined. (2) Categories can be relatively easily distinguished based on surface features in the text (e.g. particular words). (3) Researchers have sufficient domain knowledge to suggest many effective rules.

From our experience with applying a rule-based approach to categorize NPOs in a single country according to ICNPO categories based on the organizations' names, we can report that the above-mentioned criteria were fulfilled:

(1) The ICNPO provides clearly defined categories.
(2) Names of NPOs in most cases gave sufficient information to categorize the organizations according to the ICNPO. In the Austrian case, there is a legal basis for this: Association Law obliges all associations to use a name that gives an indication of their purpose and is not misleading.
(3) The research team had background knowledge about the country's nonprofit sector, and had the possibility of doing additional desk research to clarify remaining ambiguities.

The rules must be established separately for different countries, or to put it more precisely, for each language region with a specific civil society tradition – because they are based on texts that require a thorough understanding of the language and culture from which they originate.

In order to apply a rule-based approach to the Austrian case, we created dictionaries that relate words or phrases – henceforth referred to as "search terms" – in the names of the organizations to ICNPO categories. These dictionaries are arranged in tiers, like a set of sieves with ever finer meshes (see Figure 1 for an illustrative example): Each of the hierarchically arranged dictionaries is basically an IF-THEN rule. If the search term in question is part of the name of an organization that has not yet been classified, then this organization is assigned to the ICNPO category associated with the search term according to the dictionary. Hence, every rule filters some cases. The remaining uncategorized cases are subjected to the subsequently applied rules. The way the code works requires the rules to be applied sequentially, but only the sequence of rules in different tiers is important. Rules within one tier are applied in an alphabetical and hence arbitrary order.

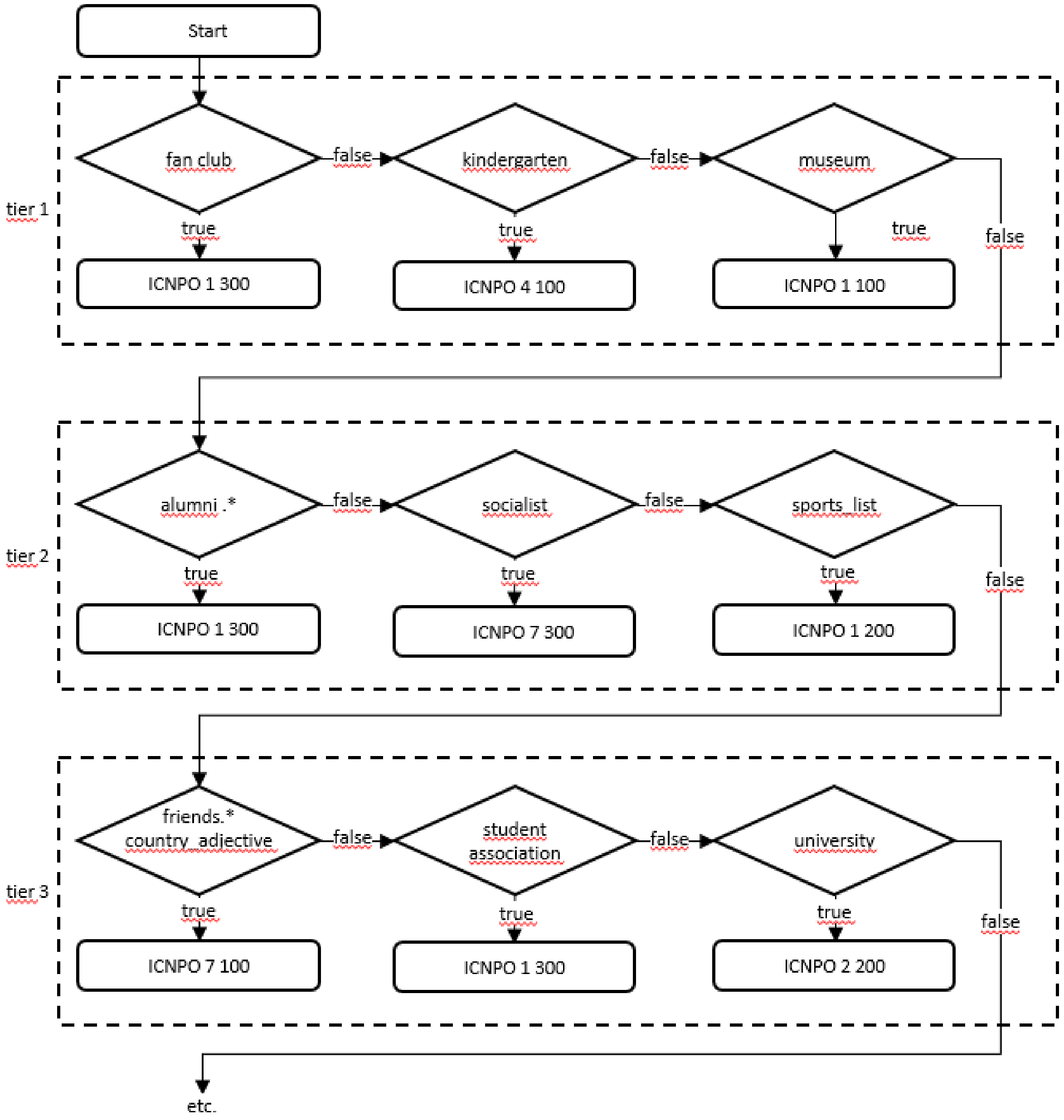


Figure 1: Example of rule-based classification in the Austrian case

A practical example is provided to illustrate this: The set of rules begins with very distinct markers, such as “fan club”, associated with the category of “other recreation and social clubs”, and “kindergarten”, associated with the category of social services. It then works its way down to increasingly hard to discern categories by progressively eliminating one ambiguity in the text after the other. E.g. when dealing with student associations, one that is named “socialist student association” goes into the category of political organizations because of the search term “socialist”, whereas remaining ones without political markers go into the category of other recreation and social clubs (see Figure 1). It must be stressed that the set of rules has to be developed inductively for each country. The above rules, for example, work only due to the fact that there are no “socialist kindergartens” or “socialist museums” in Austria.

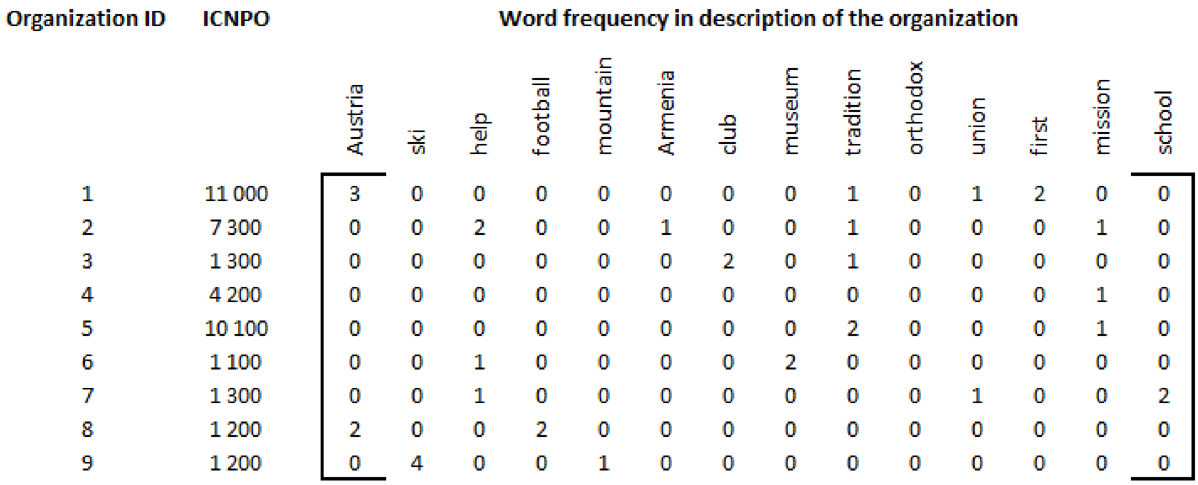
Categorization based on machine learning

The term machine learning refers to a range of approaches where the set of rules is determined through statistical procedures, as opposed to the rule-based method, where the rules are manually specified by the researcher. For machine learning, a so-called training sample is required. This is a sample of organizations for which the ICNPO category is already known. This training sample is used to train an algorithm to recognize patterns in large amounts of data that are indicative of the required piece of missing information (James, Witten, Hastie, & Tibshirani, 2017:26-28). Large amounts of text data related to the ICNPO-category of an organization are often available in genres of text that describe the organization itself or its activities. If human beings without any knowledge about human society in general could learn to categorize organizations based on these texts, then it is reasonable to assume that this task could be automated via machine learning.

As basis for the implementation of a machine learning approach in Austria we used short descriptions of the organizations retrieved from the internet[1]. These descriptions were pre-processed using natural language-processing techniques: removing stop words (i.e., words such as “and”, “or”, “this”), correcting misspellings, and stemming (i.e., reducing words to their word stem). The resulting texts were used to construct a matrix, where every word occurring in one of the descriptions represents one column (or in other words: one variable), and every row represents one organization (see Figure 2). Such a matrix is called a document term matrix. We created a training sample by drawing a sample of n=1000 organizations and manually determining the ICNPO categories of these organizations. Using this training sample, a multinomial regression model was estimated. In this model, the word columns from the document term matrix were the explanatory variables, and the ICNPO category was the dependent variable. By means of 10-fold cross validation, the regression model with the highest explanatory power was automatically identified (Kwartler, 2017:189ff.). [2]

[1] Using search engine APIs, we automatically downloaded short descriptions of websites (the so-called “snippets”) resulting from an online search for the organizations, as the algorithms applied by online search engines are most up-to-date and powerful in summarizing relevant information from text. These snippets consist of 25 to 50 words per organization.

[2] Since the document term matrix contains several thousand columns, i.e. potential explanatory variables, many of which are not expected to make a significant contribution to categorization, it is theoretically advisable to use variable selection techniques like LASSO or Ridge regression to avoid overfitting and achieve models that are more parsimonious. But neither LASSO nor Ridge models significantly improved the results in our case.



Organization ID	ICNPO	Word frequency in description of the organization													
		Austria	ski	help	football	mountain	Armenia	club	museum	tradition	orthodox	union	first	mission	school
1	11 000	3	0	0	0	0	0	0	0	1	0	1	2	0	0
2	7 300	0	0	2	0	0	1	0	0	1	0	0	0	1	0
3	1 300	0	0	0	0	0	0	2	0	1	0	0	0	0	0
4	4 200	0	0	0	0	0	0	0	0	0	0	0	0	1	0
5	10 100	0	0	0	0	0	0	0	0	2	0	0	0	1	0
6	1 100	0	0	1	0	0	0	0	2	0	0	0	0	0	0
7	1 300	0	0	1	0	0	0	0	0	0	0	1	0	0	2
8	1 200	2	0	0	2	0	0	0	0	0	0	0	0	0	0
9	1 200	0	4	0	0	1	0	0	0	0	0	0	0	0	0

Figure 2: Example of a small document term matrix.

The following example is intended to explain this basic principle in a less technical way: We gave the computer 1,000 example cases, at least 20 from every distinct ICNPO-category, together with the words associated with these examples. The computer treats these words like any other variable and estimates the coefficients of a regression model, assigning one β-coefficient to every word. Hence, every new organization's ICNPO-category can be predicted based on the words occurring in the description of the organization and the associated coefficients.

There are many different and elaborate ways of pre-processing the texts that describe the organizations (removing stop words, correcting misspellings, stemming), setting the coefficients in the document term matrix (occurrence yes/no, frequency of occurrence, normalization), and selecting the β-coefficients that perform best in predicting the desired category. Yet the principle behind machine learning is the same as in basic regression analysis: It is all about searching for correlations.

Comparing the two methods

In order to compare the performance and possibilities of rule-based categorization and categorization using machine learning, we applied both methods to a data set[3] of all non-profit associations that existed in Austria between 2006 and 2016 (n=122 514).

The machine learning approach took us 3 person-days of work to implement from scratch. It assigned an ICNPO category to every organization, but only around 20% of those categorizations were correct. The rule-based approach performed much better in terms of accuracy (see also Sokolova & Lapalme, 2009): The decision rules that we were able to set up

[3] Data was provided by Compass Verlag GmbH.

within 3 person-days of work allowed us to categorize around 40% of the organizations, with an accuracy of around 90%.

It can therefore be said that the possibilities of recognizing patterns in texts when using machine learning methods are still very limited. The state of the art is that machine-learning techniques recognize correlations between a dependent variable and words that occur in texts. This only works well if dozens of millions of cases are available for training the algorithm in the first place, which is unfortunately not the case with NPO categorization. Computers do not (yet) understand what words mean (Manning, 2016). For example: There is no algorithm that would be able to group the words “soccer” and “hockey” together, based on the commonality that both are sports. If such context information is to be taken into account, humans have to provide it in the form of additional data.

It would have been possible to improve the results of the machine-learning approach by further pre-processing the texts used to generate the independent variables. There would have been several ways to do this: Manually correcting stemming results (because we found that the stemming packages available in R for German language still leave much to be desired), creating dictionaries of synonyms or meta-categories (e.g. replacing all terms like "tennis", "swimming", etc. with "sports"), or using word sequences (so-called n-grams) as explanatory variables.

All of this would have required considerable human working time and background knowledge about the NPO sector, just as the manual creation of rules in a rules-based approach. Still, a classification based on machine learning would always have remained less transparent than a rule-based classification, because β coefficients cannot be interpreted intuitively. Their number is too large, and the coefficients in logistic regression analysis are generally difficult to interpret.

Due to the advantages of a rule-based approach in terms of accuracy and transparency, the machine-learning approach was not further pursued in the Austrian case. The human working time was used to manually formulate decision rules instead of further pre-processing the input texts for machine learning. With about 22 person-days of work, we were able to create a set of categorization rules that categorized 94% of organizations with 96% accuracy (assessed based on a manually double-checked sample of 450). The rules comprised 211 tiers, with altogether 3,090 search terms (not counting wildcat terms separately), and 10 wildcat term lists. It would have been possible to further increase the correct classification rate, but for our purposes we considered the achieved rate as sufficient. We assume that the time to specify categorization rules will be shorter for future research, as scholars can build on our algorithm, dictionaries, and recommendations for implementation.

Recommendations for implementing a rule-based approach

We implemented a rule-based approach in R (and in MS Excel for handling files of dictionaries and wildcat term lists, to be explained below). The R script as well as the dictionaries and wildcat term lists that we generated are available here under the conditions of a CC BY-NC-SA 4.0[4] license: [a link to the authors' university research depository will be provided here; for now, materials are provided through the editor to preserve the authors' anonymity].

The R script first prompts the computer to read the list of NPOs' names. These are used as the basis for categorizing the organizations according to the ICNPO. Then the script prompts reading a file with the stratified dictionaries for assigning organizations to ICNPO categories based on search terms. We also used special wildcat term lists (to be explained below) in the dictionaries. These lists are read next. Finally, as the core of the script, the dictionaries (including the wildcat term lists), are applied to categorize the organizations.

To develop the set of rules, we recommend starting with deductive coding and then shifting to inductive coding. As a starting point, it is advisable to deduce rules from the guidelines for ICNPO categorization (Salamon, 1996). The terms mentioned there should be checked for inclusion in the set of rules.

When the possibilities of identifying search terms deductively from these guidelines are exhausted, further rules need to be established inductively by examining the data. We found it helpful to use an additional short R script that calculates a term frequency matrix of all yet uncategorized organizations and to progressively work our way down from the most frequent semantically significant words.

Performance can be further improved by including wildcat term lists. These are lists of terms that are related to a particular concept. For example, the abstract concept of sport (ICNPO 1 200) is realized in many different kinds of sport. We used web scraping to obtain a list of over 200 officially recognised sports from an Austrian government website and included them in a term list to assign organisations to the ICNPO category for sport. The same approach was

[4] This means that the materials may be used and adapted for non-commercial purposes, giving credit to us as authors and sharing adapted versions under the same conditions.

taken also for names of professions and jobs, medical and health-related terms, country names, names of country citizens and ethnic groups, various kinds of animals, as well as towns and regions within Austria. These wildcat term lists can be included in the dictionary rules like variables. For example, in the search term “friends.* country_adjective”, the term “country_adjective” serves as a wildcat for the full list of countries in their adjective form (e.g. Armenian, Chinese…). With the use of such wildcat terms and search modifiers (especially .* as a placeholder for a flexible number of characters) it is possible to build up an elaborate and precise system of categorization rules.

Each suspected new rule should be cross-checked with the full sample: It should be included if it returns no (or only a negligible[5] number of) false hits. To facilitate cross-checking, we recommend using a preliminary ICNPO marker that includes the tier on which the organization was categorized and generously adding new tiers. If necessary, additional tiers can be added in retrospect by re-assigning tier numbers with decimal places.

Conclusion

Applying two different computational methods for categorizing nonprofit organizations, we find that building up a tool for valid categorization is indeed feasible with widely available software. Where good quality administrative data are not available, computational categorization methods may be an acceptable alternative for acquiring the desired information.

We recommend a rule-based approach to computational categorization because machine-learning approaches are not (yet) able to achieve high-precision results without much human effort for pre-processing text data. In addition, classifications based on machine learning are relatively non-transparent, because they are based on statistical models that are too large and too complex for intuitive interpretation.

Rule-based approaches quickly produce relatively precise results, and they are completely transparent because decision-making rules are explicitly written down and not hidden in a black box of statistical models. However, it must be acknowledged as a major disadvantage of rule-based approaches that developing the set of rules – although sometimes surprising and

[5] Some categorization error might be traded for higher overall categorization rates. If e.g. one is interested in mapping a country’s NPO-sector as a whole, rules yielding small false-positive but high true-positive rates might be considered for inclusion.

amusing – requires thorough contextual knowledge about the nonprofit sector in the place of interest and is not one of the most exciting and enjoyable kinds of research labor.

References

Grønbjerg, K. A., & Paarlberg, L. (2002). Extent and nature of overlap between listings of IRS tax-exempt registration and nonprofit incorporation: The case of Indiana. *Nonprofit and Voluntary Sector Quarterly, 31*(4), 565-594.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. New York, NY et al.: Springer.

Kwartler, T. (2017). *Text mining in practice with R*. Hoboken, NJ: Wiley.

Lecy, J., & Thornton, J. (2016). What Big Data can tell us about government awards to the nonprofit sector. *Nonprofit and Voluntary Sector Quarterly, 45*(5), 1052-1069.

Manning, C. (2016). Understanding human language: Can NLP and Deep Learning help? *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*.

McDonnell, D., & Rutherford, A. C. (2018). The determinants of charity misconduct. *Nonprofit and Voluntary Sector Quarterly*, *47*(1), 107-125.

Salamon, L. M., & Anheier, H. K. (1992). In search of the non-profit sector II: The problem of classification. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *3*(3), 267-309.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Zhai, C. X., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining*. New York, NY: Association for Computing Machinery and Morgan & Claypool.