HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jides

Mining online political opinion surveys for suspect entries: An interdisciplinary comparison

Costantinos Djouvas[a,*], *Fernando Mendez*[b], *Nicolas Tsapatsoulis*[a]

[a] *Cyprus University of Technology, Department of Communication and Internet Studies, 94 Anexartisias St. Iakovides building, 3rd Floor, Limassol 3040, Cyprus*

[b] *Electronic Democracy Centre (e-DC), Zentrum für Demokratie Aarau ZDA, Kuttigerstr. 21 CH - 5000 Aarau, Switzerland*

ARTICLE INFO

Article history:
Published online 2 December 2016

Keywords:
Voting advice applications
Data cleaning
Machine learning
Data mining
Anomaly detection
Psychometric Likert scale

ABSTRACT

Filtering data generated by so-called Voting Advice Applications (VAAs) in order to remove entries that exhibit unrealistic behavior (i.e., cannot correspond to a real political view) is of primary importance. If such entries are significantly present in VAA generated datasets, they can render conclusions drawn from VAA data analysis invalid. In this work we investigate approaches that can be used for automating the process of identifying entries that appear to be suspicious in terms of a users' answer patterns. We utilize two unsupervised data mining techniques and compare their performance against a well established psychometric approach. Our results suggest that the performance of data mining approaches is comparable to those drawing on psychometric theory with a fraction of the complexity. More specifically, our simulations show that data mining techniques as well as psychometric approaches can be used to identify truly 'rogue' data (i.e., completely random data injected into the dataset under investigation). However, when analysing real datasets the performance of all approaches dropped considerably. This suggests that 'suspect' entries are neither random nor clustered. This finding poses some limitations on the use of unsupervised techniques, suggesting that the latter can only complement rather than substitute existing methods to identifying suspicious entries.

1. Introduction

In this paper we draw on data generated by an EU-wide Voting Advice Application (VAA), called EUvox. VAAs are freely available web tools that match the preferences of voters to that of candidates or political parties [1–5]. The mechanism of the VAA is simple. A set of experts compile a collection of important issues or policy statements, $Q = \{Q_1, Q_2, \ldots, Q_k\}$, upon which users have to express their opinion by selecting one of several categories (or answers). In the majority of cases the number of policy statements (referred to also as questions or items) is 30. Furthermore,

Peer review under responsibility of Qassim University.

* *Corresponding author.*
E-mail addresses: costas.tziouvas@cut.ac.cy (C. Djouvas), fernando.mendez@zda.uzh.ch (F. Mendez), nicolas.tsapatsoulis@cut.ac.cy (N. Tsapatsoulis).

http://dx.doi.org/10.1016/j.jides.2016.11.003

the possible answers are defined as a Likert scale with the following options: 'Strongly Disagree', 'Disagree', 'Neither Agree nor Disagree', 'Agree', 'Strongly Agree', as well as a 'No opinion' category that are encoded as follows: 1 = 'Strongly Disagree', 2 = 'Disagree', 3 = 'Neither Agree nor Disagree', 4 = 'Agree', 5 = 'Strongly Agree', and 6 = 'No opinion'. Thus, every answer can take values in the set $A = \{1, 2, 3, 4, 5, 6\}$. Moreover, we define S_i to be the sequence of answers of the ith user to the policy statement (e.g. $S_i = 2, 4, 3, \ldots, 1, 6, 5$), where $|S_i|$ = number of policy statements. Let us denote with a_i^j to be the jth element ($j \in \{1, 2, \ldots, |S_i|\}$) of the ith sequence (typically in a VAA setting this sequence corresponds to the filled in questionnaire of the ith VAA user).

In the absence of any prior information the probability to get any of the values in A is uniform, i.e.,:

$$p(a_i^j = k) = \frac{1}{|A|}, \tag{1}$$

where $k \in A$. In practice, however, because the sequences are generated by voters that express a finite (actually a small) number of political views some answers are more probable than others. For instance, the positive and negative answers in the middle categories of the scale (i.e., Disagree, and Agree) are more probable than the extreme ones (i.e., Strongly Disagree, and Strongly Agree); this is due to the tendency of humans to avoid taking extreme positions in political statements such as those included in a VAA [6]. Furthermore, the meaning of a 'No opinion' response category attracts itself a special research interest, while its probability of appearance is in general lower than other response options [6]. Fig. 1 depicts the distribution of responses for each of the 30 policy statements for one of the datasets – Ireland – that is used in this paper.[1]

Thus far, we have not taken into consideration the content of the policy statements. In a real setting, however, where users are completing a VAA questionnaire to receive a vote recommendation, there is a highly influential factor that conditions the selection of answers to policy questions: the political view or ideology of a user. It is precisely for this reason that analysts have used VAA generated data to study the dimensionality of the political space by identifying configurations of latent dimensions, such as a left–right or liberal–conservative [7–10]. The fact that most users have a political ideology ensures that there is some degree of consistency in answer responses. For instance, if a user has right-wing political preferences for policy A, B and C they are also likely to have a right-wing political preference for policy D. Evidently, this assumes that A, B, C and D hang together in the policy space. VAA designers have some a-priori knowledge about established dimensions of political competition and design their questionnaires to include groups of policy items that are related to each other [11,9,10]. For instance, the EUvox had three categories related to (1) Europe, (2) economy and (3) broader societal issues. Most users' response patterns would therefore entail some degree of ideological consistency across these dimensions. This has important implications since it makes some sequences of answering patterns more probable than others, which renders any assumption concerning the independence of VAA policy statements invalid. In a mathematically rigorous way this can be stated as follows: let $S = \{S_1, S_2, \ldots, S_n, \}$ be a set of sequences produced by n VAA users. The conditional probability $P(S_i|S)$ is very different than the probability $P(S_i)$. If we assume independence between the VAA statements then $P(S_i)$ and $P(S_i|S)$ can be computed as follows:

$$P(S_i) = \prod_{j=1}^{n} p(a_i^j) \tag{2}$$

and

$$P(S_i|S) = \prod_{j=1}^{n} p(a_i^j|S). \tag{3}$$

If a 'legitimate' VAA user is likely to answer the questionnaire in an ideologically structured manner, and this applies to most users attracted to a VAA, then randomly and inconsistently generated sequences are likely to be quite rare in a real VAA setting. If such sequences in S exist, then they would constitute suspicious entries and should be removed prior to performing any data analysis. We know from various studies that VAA data do contain rogue entries although to our knowledge no filtering is undertaken using structured pattern analyses of responses (see [12,13] for a review). Current best practices [14,3,15–17,13] remove entries by collecting and utilizing some para-data (such as total time taken to complete questionnaire and IP filters). These have sometimes been complemented with the removal of entries with dubious answer patterns, such as too many 'No opinions' or many consecutive same answer responses. Such filters are easy to implement and do not require any sophisticated analysis—though they can become quite arbitrary. If researchers are serious about their cleaning methods, a powerful case has been made for the need to collect data based on individual item response timers [12]. As we shall see below, this paper builds on this key insight.

The identification of suspect entries in questionnaires is a field of study within psychometrics (see [18] for a review). Specifically, there is a literature derived from Item Response Theory that focuses on identifying such inconsistencies in test scores such as exams [19,20]. In data mining terms these inconsistencies are usually described as anomalous and there exists within the field a number of unsupervised anomaly detection techniques that are applied in different areas and disciplines. For example unsupervised anomaly detection can be used for intrusion detection [21,22], for fraud detection [23,24], in medicine for disease detection [25] and for prescription control [26], in image processing for identifying foreign object [27,28], and in text data for novelty detection [29]. However, to the best of our knowledge, there is very limited literature on applying data mining techniques to data consisting of responses to Likert items, as is the case in VAA questionnaires. [30] applied a number of both supervised and unsupervised techniques for extracting patterns in students' evaluations of their instructors where Davier in [31] used data mining techniques to address the problem of testing quality control.

[1] See Section 3 for more information about the datasets used.