Subgroup Discovery for Election Analysis: A Case Study in Descriptive Data Mining

Henrik Grosskreutz, Mario Boley, and Maike Krause-Traudes

Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany
{firstname.lastname}@iais.fraunhofer.de

Abstract. In this paper, we investigate the application of descriptive data mining techniques, namely subgroup discovery, for the purpose of the ad-hoc analysis of election results. Our inquiry is based on the 2009 German federal Bundestag election (restricted to the City of Cologne) and additional socio-economic information about Cologne's polling districts. The task is to describe relations between socio-economic variables and the votes in order to summarize interesting aspects of the voting behavior. Motivated by the specific challenges of election data analysis we propose novel quality functions and visualizations for subgroup discovery.

1 Introduction

After a major election of public interest is held, there is a large and diverse set of societal players that publishes a first analysis of the results within the first day after the ballots are closed. Examples include traditional mass media like newspapers and television, citizen media like political blogs, but also political parties and public agencies. An instance of the last type is the Office of City Development and Statistics of the City of Cologne. The morning after major elections that include Cologne's municipal area, the office publishes a first analysis report on the results within the city[1]. In this report, socio-economic variables (e.g., average income, age structure, and denomination) are related to the voting behavior on the level of polling districts. The Office of City Development and Statistics performs much of the analysis, such as selecting a few candidate hypotheses, beforehand, i.e., based on previous election results—a course of action that might neglect interesting emerging developments. However, due to the strict time limit involved, there appears to be no alternative as long as an analyst mainly relies on time-consuming manual data operations. This motivates the application of semi-automatized data analysis tools.

Therefore, in this academic study, we take on the perspective of an analyst who is involved in the publication of a short-term initial analysis of election

[1] The report on the 2009 Bundestag election can be found (in German language) at http://www.stadt-koeln.de/mediaasset/content/pdf32/wahlen/bundestags\wahl2009/kurzanalyse.pdf

B. Pfahringer, G. Holmes, and A. Hoffmann (Eds.): DS 2010, LNAI 6332, pp. 57–71, 2010.

Party	description	result 2009	change
SPD	social democrats	26%	-12.1
CDU	conservatives	26.9%	-0.3
FDP	liberals	15.5%	+4
GRUENE	greens	17.7%	+2.8
LINKE	dem. socialists	9.1%	+3.4

Fig. 1. Results of the 2009 Bundestag election in Cologne

results, and we investigate how data mining can support the corresponding ad-hoc data analysis. In order to narrow down the task, we focus on the following *analysis question*:

> *What socio-economic variables characterize a voting behavior that considerably differs from the global voting behavior?*

This question is of central interest because it asks for interesting phenomenons that are *not* captured by the global election result, which can be considered as base knowledge in our context. Thus, answers to this question have the potential to constitute novel, hence, particularly news-worthy, knowledge and hypotheses. This scenario is a prototypic example for *descriptive* knowledge discovery: instead of deducing a global data model from a limited data sample, we aim to discover, describe, and communicate interesting aspects of it.

(a) (b)

Fig. 2. Spatial visualization of polling districts. Color indicates: (a) above average FDP votes; (b) high share of households with monthly income greater than 4500€

We base our study on the German 2009 federal Bundestag election restricted to the results of Cologne. For this election we analyzed the data during a corresponding project with Cologne's Office of City Development and Statistics. See Figure 1 for the list of participating parties, their 2009 election results, and the difference in percentage points to their 2005 results. The data describes the election results on the level of the 800 polling districts of the city, i.e., there is one data record for each district, each of which corresponds to exactly one polling place. Moreover, for each district it contains the values of 80 socio-economic variables (see Appendix A for more details). Figure 2 shows the geographical