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.

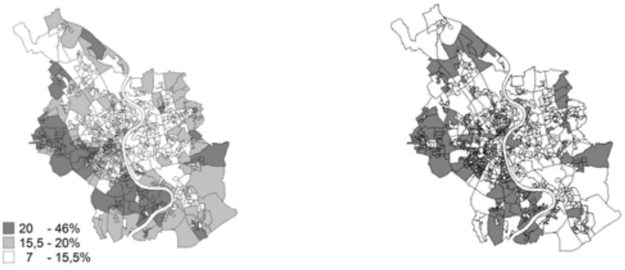
Party	description	result 2009	change
🟥 SPD	social democrats	26%	-12.1
⬛ CDU	conservatives	26.9%	-0.3
🟨 FDP	liberals	15.5%	+4
🟩 GRUENE	greens	17.7%	+2.8
🟪 LINKE	dem. socialists	9.1%	+3.4

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 alignment of the districts. To illustrate that indeed there is a relation between voting behavior and socio-economic variables, the figure additionally shows the districts with a high share of households with high income (Fig. 2(b)). The high overlap between these districts and those with above-average votes for party FDP (Fig. 2(a)) is an indication that those two properties are correlated.

The semi-automatized data analysis we propose in this work, i.e., the descriptive pattern and hypotheses generation, is meant purely indicative: the evaluation of all discovered patterns with respect to plausibility (e.g., to avoid ecological inference fallacy [18]) and their interestingness are up to the analyst and her background knowledge (hence, *semi-automatized* data analysis). In fact, almost all of the analyst's limited time and attention has to be reserved for the manual preparation and creation of communicable content. Thus, in addition to fast execution times, a feasible tool has to meet the following requirements:

(R1) The tool has to discover findings that directly support answers to the analysis question above. In particular, it must suitably define how to assess "notably different voting behavior." That is, it must provide an *operationalization* that relates this notion to the measurement constituted by the poll.

(R2) Operating the tool has to be simple. In particular, it either has to avoid complicated iterative schemes and parameter specifications, or there must be clear guidelines for how to use all degrees of freedom.

(R3) The tool's output must be intuitively interpretable and communicable. This involves avoiding redundant or otherwise distracting output as well as providing a suitable visualization.

In the next section we show how *subgroup discovery* can be configured to meet these requirements. In particular, we propose new quality functions for subgroup discovery that are capable of handling *vector-valued target variables* as they are constituted by election results, and that *avoid the generation of redundant subgroups* without requiring the specification of any additional parameter. Moreover, we propose a *visualization technique* that is tailor-made for rendering subgroups with respect to election results.

2 Approach

There are existing approaches to election analysis (e.g., [8,14]) that are either based on a global regression model or on an unsupervised clustering of the population. In contrast, we approach the problem from a supervised local pattern discovery perspective [15]. As we show in this section, our analysis question naturally relates to the task of *subgroup discovery* [21]. After a brief introduction to this technique, we discuss how a subgroup discovery system can be configured and extended such that it satisfies the initially identified requirements (R1)-(R3).

2.1 Subgroup Discovery

Subgroup discovery is a descriptive data mining technique from the family of *supervised descriptive rule induction methods* (see [17,16]; other members of the family include *contrast set mining*, *emerging patterns* and *correlated itemset mining*). It aims to discover local sub-portions of a given population that are a) large enough to be relevant and that b) exhibit a substantially differing behavior from that of the global population. This difference is defined with respect to a designated target variable, which in our scenario is the election result. The data sub-portions are called *subgroups*, and they are sets of data records that can be described by a conjunction of required features (in our case the features are constraints on the values of the socio-economic variables).

For a formal definition of subgroup discovery, let DB denote the given database of N data records $d_1, \ldots, d_N$ described by a set of n (binary) features $(f_1(d_i), \ldots, f_n(d_i)) \in \{0,1\}^n$ for $i \leq N$. A *subgroup description* is a subset of the feature set $sd \subseteq \{f_1, \ldots, f_n\}$, and a data record d satisfies sd if $f(d) = 1$ for all $f \in sd$, i.e. a subgroup description is interpreted conjunctively. The *subgroup* described by sd in a database DB, denoted by $DB[sd]$, is the set of records $d \in DB$ that satisfy sd. Sometimes, $DB[sd]$ is also called the *extension* of sd in DB. The interestingness of a subgroup description sd in the context of a database DB is then measured by a *quality function* q that assigns a real-valued quality $q(sd, DB) \in \mathbb{R}$ to sd. This is usually a combination of the subgroup's size and its unusualness with respect to a designated target variable.

In case the target variable T is real-valued, that is, T is a mapping from the data records to the reals, the unusualness can for instance be defined as the deviation of the mean value of T within the subgroup from the global mean value of T. A common choice is the *mean test quality function* [9]:

$$q_{mt}(DB, sd) = \sqrt{\frac{|DB[sd]|}{|DB|}} \cdot (m(DB[sd]) - m(DB)) \tag{1}$$

where $m(D)$ denotes the mean value of T among a set of data records D, i.e., $m(D) = 1/|D| \sum_{d \in D} T(d)$.

Generally, quality functions order the subgroup descriptions according to their interestingness (greater qualities correspond to more interesting subgroups). One is then usually interested in k highest quality subgroup descriptions of length at most l where the length of a subgroup description is defined as the number of features it contains.

2.2 Application to Election Analysis

Requirement (R1) of Section 1 includes the support of a suitable operationalization of “notably different voting behavior”. If we choose to perform this operationalization on the level of individual parties, i.e., as a notably different share of votes of one specific party, the analysis question from Section 1 can directly be translated into subgroup discovery tasks: choose q_{mt} as quality function, create a set of features based on the socio-economic variables, and as target variable T choose either a) the 2009 election result of a particular party or b) the difference of the 2009 and the 2005 result. The latter option defines “voting behavior” with respect to the *change* in the share of votes. This is a common perspective that is usually used to interpret the outcome with respect to the success or failure of individual parties.

In order to answer the analysis question independently of a specific party, one can just run these subgroup discoveries once for each of the possible party targets and then choose the k best findings among all returned patterns. With this approach, the patterns for the overall voting behavior are chosen from the union of the most interesting patterns with respect to the individual parties.

There are, however, several important relations among the parties. For instance, they can be grouped according to their ideology (in our case, e.g., SPD, GRUENE, and LINKE as "center-left to left-wing"), or one can distinguish between major parties (SPD and CDU) and minor parties (FDP, GRUENE and LINKE). Voting behavior can alternatively be characterized with respect to such groups (e.g., "in districts with a high number of social security claimants, the major parties lost more than average.") This indicates that a subgroup may be interesting although no individual party has an interesting result deviation in it, but because the *total* share of two or more parties is notably different.

This is not reflected in the initial approach, hence it is desirable to extend the subgroup discovery approach such that it captures requirement (R1) more adequately. In particular, we need a quality function that does not only rely on a single target variable but instead on a set of k real-valued target variables $T_1, \ldots, T_k$. With this prerequisite we can define a new quality function analogously to the mean test quality by

$$q_{\mathrm{dst}}(DB, sd) = \sqrt{\frac{|DB[sd]|}{|DB|}}\, \|\mathbf{m}(DB[sd]) - \mathbf{m}(DB)\|_1 \tag{2}$$

where $\mathbf{m}(D)$ denotes the mean vector of the $T_1, \ldots, T_k$ values among a set of data records D, i.e.,

$$\mathbf{m}(D) = 1/|D| \sum_{d \in D} (T_1(d), \ldots, T_k(d)) \ .$$

and $\|(x_1, \ldots, x_k)\|_1$ denotes the 1-norm, i.e., $\sum_{i=1}^{k} |x_i|$. Using this quality function we arrive at an alternative instantiation of subgroup discovery. We can choose q_{dst} as quality function and either a) T_i as the 2009 share of votes of party i or b) T_i as the gain (2009 result minus 2005 result) of that party.

2.3 Avoidance of Redundant Output

Requirement (R3) demands the avoidance of redundant output, but, unfortunately, a problem with the straightforward discovery of subgroups and other descriptive patterns is that a substantial part of the discovered patterns can be very similar. That is, many patterns tend to be only slight variations of each other, essentially describing the same data records. The reason for this is twofold. Firstly, there may be many highly correlated variables that provide interchangeable descriptions. We can get rid of these by performing a correlation analysis during preprocessing (see Section 3.1). In addition, for an interesting subgroup *sd* it is likely that there are some strict *specializations* $sd' \supset sd$ with an equal or slightly higher quality. Although the truly relevant and interesting portion of the subgroup may be described most adequately by *sd*, those specializations are at least equally likely to appear in the output, causing redundancy or—even worse—pushing *sd* out of the result set altogether.

We now present an approach that generalizes a common principle of some of the existing methods to address this problem [3,7,19,20], namely to discard subgroups *sd* that do not substantially improve their strict *generalization* $sd' \subset sd$. As captured in our requirement (R2) we want to avoid the introduction of additional parameters. Therefore, unlike the cited approaches, we do not introduce a minimum improvement threshold, but instead we *use the quality function itself* to measure the sufficiency of an improvement. That is, for some arbitrary *base* quality function q, we propose to assess the quality of a pattern *sd* as the minimum of the quality of *sd* with respect to the extension of all its generalizations. More precisely, we consider the quality function q^{Δ} that is defined as $q^{\Delta}(DB, \emptyset) = q(DB, \emptyset)$ for the empty subgroup description $\emptyset$ and

$$q^{\Delta}(DB, sd) = \min_{sd' | sd' \subset sd} q(DB[sd'], sd) \tag{3}$$

otherwise. We call q^{Δ} the *incremental version* of q. After giving some additional definitions, we discuss in the remainder of this subsection that q^{Δ} has some desirable properties.

We call a subgroup description *sd tautological* with respect to a database *DB* if $DB[sd] = DB$, and we call *sd non-minimal* with respect to *DB* if there is a generalization $sd' \subset sd$ having the same extension, i.e., $DB[sd] = DB[sd']$. Moreover, we say that a quality function q is *reasonable* if $q(DB, sd) \leq 0$ whenever *sd* is tautological with respect to *DB*.

Proposition 1. *Let DB be a database, q a quality function, and* q^{Δ} *its incremental version. If q is reasonable, then* q^{Δ} *is non-positive for all non-minimal subgroup descriptions in DB.*

Proof. Note that, by definition, every non-minimal subgroup has a strict generalization sd' with identical extension. Therefore,

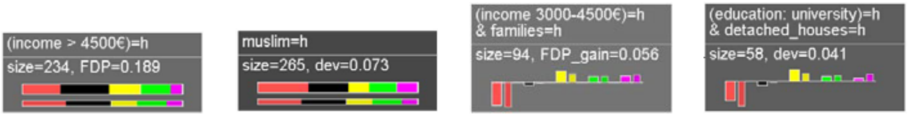
$$q^{\Delta}(DB, sd) \leq q(DB[sd'], sd) = q(DB[sd], sd) = 0$$

where the last equality follows from q being reasonable. □

This property assures that non-minimal subgroup descriptions are filtered from the result set. Such descriptions are considered redundant respectively trivial [4,19]. For quality functions based on the mean deviation (e.g., Eq. 1) an even stronger statement holds: for such functions all descriptions are filtered that do not provide an improvement in the mean deviation. Thus, the incremental quality directly follows other filtering paradigms from descriptive rule induction; namely it eliminates patterns that do not provide a *confidence improvement* [3] respectively that are not *productive* [20]. Finally, we remark that the incremental quality is bounded by the base quality, i.e., for all subgroup descriptions sd it holds that $q^{\Delta}(DB, sd) \leq q(DB, sd)$.

2.4 Visualization

In order to completely meet the last requirement (R3), we need an appropriate visualization technique. Although there is existing work on subgroup visualization (e.g., [1,10]), we choose to design a new technique that is tailor-made for election analysis and allows for multiple target attributes. In fact we propose four visualizations, one for each possible subgroup discovery configuration discussed in Section 2.2. A common element is that every subgroup is visualized by a grey box having a color intensity that reflects the subgroup's quality. Higher qualities correspond to more intense grey shades. Every box shows the subgroup description and the size of its extension, plus additional information that depends on the quality function as well as on the operationalization of "voting behavior".



(a) (b) (c) (d)

Fig. 3. Visualization for the different combinations of quality functions and operationalizations of election result: (a) single party result, (b) result distribution, (c) single party gain and (d) gain/loss vector. The result of the particular parties are plotted using their official colors, listed in Figure 1.

Figure 3 shows the four cases: (a) absolute results of a single party, (b) combined absolute results for all parties, (c) gain for a single party, and (d) combined gains and losses of all parties. In case the mean test is used as quality function, we show the mean value of the target variable, i.e. the result for a particular party, in the subgroup next to the extension size. For the vector-valued quality function this space is occupied by the 1-norm of the mean vector difference. Beside this figure, the boxes include a visualization of the election result of all parties. Depending on whether absolute 2009 results or the gains with respect to 2005 are considered, the results are rendered in a different fashion. The absolute 2009 results are represented by two bars (Figure 3(a) and 3(b)): the upper bar corresponds to the distribution over parties in the subgroup, while the lower bar visualizes the overall distribution. The different segments in the bars represent the share of votes for the different parties. They are visualized from left to right using the parties' official colors: red (SPD), black (CDU), yellow (FDP), green (GRUENE) and magenta (LINKE). If, instead, the gains respectively losses are considered (Figure 3(c) and 3(d)), the result is displayed as bar chart that is centered around a gain of 0. Gains are visualized by upward bars, while losses are visualized by downward bars. Again, the global gains are also plotted for easy comparison: For every party, a first bar shows the local gains in the subgroup, while a second bar on its right-hand shows the global gains. This second bar provides the context information required to interpret the gains in a particular subgroup.

3 Experiment

After the introduction of our tools we are now ready to describe our case study on the 2009 Bundestag election. Before we provide and discuss the results we briefly summarize our experimental setup.

3.1 Setup

From the raw input data to the final output we performed the following steps.

preprocessing. In order to avoid the occurrence of highly correlated features in the result, we performed a correlation analysis and removed one variable out of every pair of variables with a correlation of at least 0.85. The choice was based on background-knowledge and subjective preference. Moreover, we performed a 3-bin frequency discretization to all remaining numerical variables.

features. Based on the discretization, we defined the set of descriptive features as follows: for every variable and every bin, there is a binary feature that a data record possesses if and only if the variable value of this record lies in that bin. These features are denoted $\mathtt{V}=\mathtt{h}$, $\mathtt{V}=\mathtt{m}$, and $\mathtt{V}=\mathtt{l}$, respectively. There are, however, several exceptions. Some variables are part of a set of complementary variables that together describe a common underlying measurement. For instance, for the age structure there is one variable representing the number of inhabitants aged 16-24, the inhabitants aged 25-34, and so on, respectively. For such variables, we did not create features corresponding to the middle or lower bin because they would have only low descriptive potential. Altogether, there is a total of 64 descriptive features.

parameters. We used a length limit l of 3 for the subgroup descriptions, and a number of subgroups k of 10. These settings lead to a reasonably small set of results that can be manually inspected and that are short enough to be easily communicable.

targets. As stated in Section 2.2 there are several options for the operationalization of "voting behavior": one has to choose between individual parties and the combined results as well as between the absolute (2009) results and the difference between the 2009 and the previous (2005) results. This leaves us with four different configurations of quality functions and target variables.

C1. For absolute combined results, quality function $q = q^{\Delta}_{\mathrm{dst}}$ (Eq. 2) with target variables $T_1, \ldots, T_5$ such that T_i is the 2009 share of votes of party i.

C2. As exemplary configuration for absolute results in the single party case, $q = q^{\Delta}_{mt}$ (Eq. 1) with the 2009 result of FDP as target T.

C3. For combined results measured by the difference to previous elections, $q = q^{\Delta}_{\mathrm{dst}}$ with target variables $T_1, \ldots, T_5$ such that T_i as the gain (2009 result minus 2005 result) of party i.

C4. Again as exemplary configuration for differences in the single party case, $q = q^{\Delta}_{mt}$ with the difference between the 2009 and the 2005 result of FDP as target.

visualization. Finally, for each configuration the resulting subgroups are rendered using the appropriate visualization technique introduced in Section 2.4. Additionally, the boxes are joined by arrows corresponding to the transitive reduction of the specialization relation among the subgroups.

Some of the above steps are not fully consistent with our requirement (R2). In particular, in the preprocessing step the user is left with the decision which variables to keep. Moreover, the parameter settings (for the number of bins and the number of subgroups) are not the only viable option. However, they are a good starting point, given that a restriction to 3 bins results in bins with an easily communicable meaning ("low" and "high"), while 10 subgroups represent a manageable amount of patterns.

3.2 Results

After describing the setup of our experiments we now present the result it yielded. In order to put our findings into context, we first recap the most important aspects of the 2009 Bundestag election results: The parliamentary majority shifted from the so-called grand coalition (CDU and SPD) to a coalition of CDU and FDP. The expiration of the grand coalition was essentially caused by an all-time low result of the social-democratic SPD combined with substantial gains for the FDP. This development is also reflected in the local results of Cologne (see Figure 1).

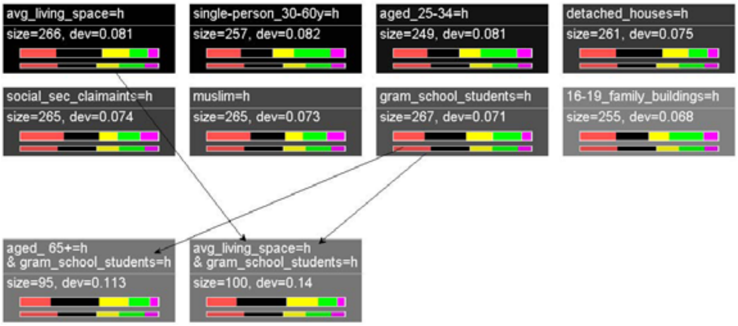


Fig. 4. Subgroups found using the distribution over parties as label

Absolute results of all parties. Figure 4 shows the subgroups obtained using Configuration C1, i.e., considering the combined absolute results of all parties in the 2009 election. There are several subgroups with a strong preference for the liberal-conservative election winners, FDP and CDU. These include the subgroup of districts with a "high average living space per accommodation," and the subgroup "high share of detached houses." The longer subgroup description "high average living space per accommodation and high share of grammar school students" is even more notable, as it has an extremely high share of FDP votes. While all other parties have lower results in these subgroups, the share of LINKE votes is particularly low. Another interesting subgroup is "high number of 30-60 year-old single-persons." This constraint is an indicator for a high share of GRUENE voters. All other parties obtained results below average in this subgroup, those of the CDU being particularly weak. There are also subgroups with a high share of SPD and LINKE votes, namely "high share of social security claimants" and "high share of muslims."

This first experiment shows that our tool reveals features that imply a strong voting preference for one particular party (e.g., GRUENE) as well as for political alliances or ideological blocks (e.g., CDU/FDP and SPD/LINKE). It is important to note that subgroups of the latter kind—although they have a clear interpretation and are easily communicable—can be missed if the analysis is performed using the single party operationalization: if one uses this option, for instance the "high share of social security claimants" subgroup is not among the top-10 subgroups.

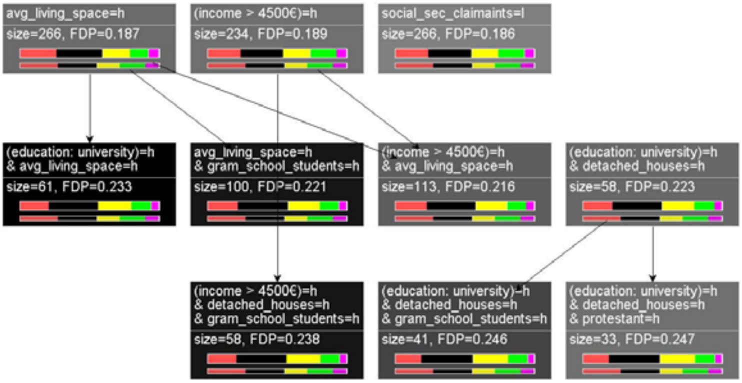


Fig. 5. Subgroups for target 'FDP'

Absolute result of FDP. Still, in case one is solely interested in one particular party, it is a reasonable choice to resort to the individual party configurations. Configuration C2 exemplary considers the FDP 2009 results—the party with the highest gain. Figure 5 shows corresponding subgroups.

While the figure shows some subgroups which are already identified using Configuration C1 (e.g., “high average living space” and the specialization with the additional constraint “high share of grammar school students”), it also contains additional results. For instance, districts with a high share of persons with “net income of more than 4500€” (see Figure 2(b) for a geographical visualization). This feature is confirmed by many other investigations to be an attribute associated with FDP voters.

Gains of all parties. We now move on to the alternative operationalization of voting behavior based on the gains respectively losses. Again, first we consider the combined gains and losses of all parties as specified in Configuration C3. Figure 6 shows the result.

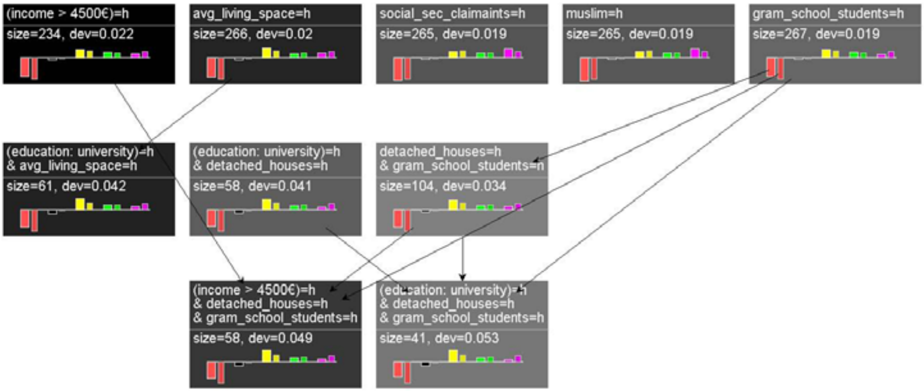


Fig. 6. Subgroups found using the distribution over the gains as label

The districts with a high share of persons with a “net income over 4500€” experienced over-average gains for the FDP, as well as (small) losses for the CDU. Such slightly over-average CDU losses can also be observed in the other subgroups with very strong FDP gains, like “average living space per accommodation” or the longer description “detached houses, grammar school students and high income”. Another interesting observation is that these subgroups are also considered in the previous section, in which we considered subgroups with a high absolute share of FDP votes. This co-ocurrence indicates that FDP could achieve additional gains in its party stronghold. The inverse relation can also be observed for the SPD subgroups: the districts with a high share of “social security claimants”—which were observed to have high SPD results—actually witnessed above-average SPD losses. The same holds for the districts with a high share of muslims. This observation suggests that the SPD is losing popularity right in its party strongholds; an assumption shared by a broad range of media analysts.

One advantage of the visualization is that it not only allows identifying the winners in a subgroup, but that it also indicates where the votes could have come from. In the two subgroups above, which attract attention due to over-average SPD results *and* over-average SPD losses, the clear winner is the LINKE, while none of the other parties have above-average gains. This is a hint that a large part of a former SPD stronghold turned into LINKE voters.

Gains of FDP. Finally, it is possible to search for subgroups with particular gains or losses of a particular party. Using Configuration C4 we exemplarily do so again with the FDP gains. The result is shown in Figure 7. Beside confirming some results from the all parties configuration, it also reveals some additional observations. The most noteworthy is perhaps the subgroup with a high share of families having an upper middle-class monthly income (i.e. 3000-4500€). This group is not traditionally associated with the FDP, and can thus constitute a hypothetical part of an explanation of the FDP's success in this election.

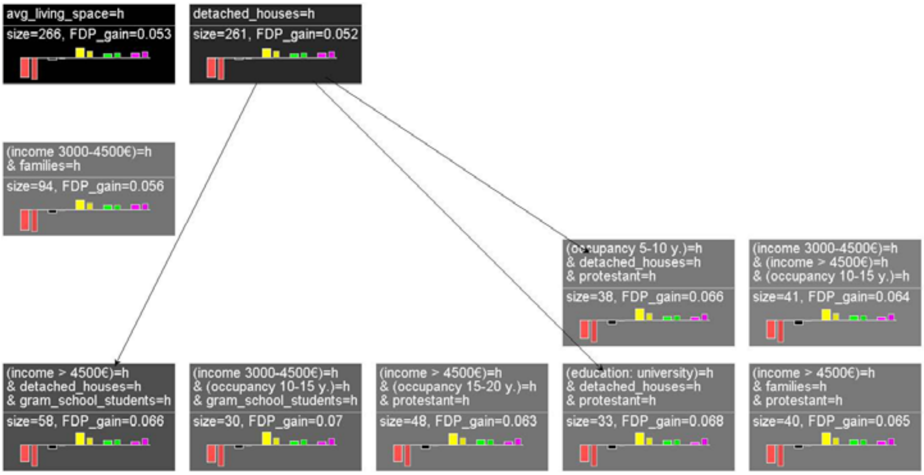


Fig. 7. Subgroups with high FDP gain

3.3 Comparison with the Traditional Approach

It is interesting to compare the results presented here with the findings reported in the Cologne report mentioned in the introduction. The main question considered there is the identification of party hot-spots and their characterization by socio-demographics attributes. This corresponds to our analysis configuration C2, which considers the absolute result of a particular party. If we compare our results with the report, we observe that our algorithm reveals the same socio-economic variables as those selected by the Cologne experts in a time-consuming manual investigation based on prior knowledge and experience. In the case of FDP, for example, the Cologne report also selects the proportion of persons with a high income and the grammar school students ratio to characterize polling districts with a high FDP support (see page 34 of the report).

3.4 Scalability

While the (manual) preprocessing steps can require some time depending on the complexity of the given data, the actual subgroup discovery is fast: for each of the four configurations, the computation takes less than 30 seconds on a standard Core 2 Duo E8400 PC. A detailed analysis of complexity issues is beyond the scope of this paper, but we not that subgroup discovery scales well in practice [2,6]—in particular with the numbers of polling districts, which is the quantity that is expected to vary the most in case the method is applied to other elections. Hence, given that the preprocessing is done in advance, the approach can be applied, e.g., during an election night.

4 Summary and Discussion

In this paper, we have demonstrated the application of a descriptive data mining technique, namely subgroup discovery, to ad-hoc election data analysis. This demonstration included a case study based on the 2009 Bundestag elections restricted to the data of Cologne. Besides presenting the results of this study, we formulated several requirements for data analysis software in this application context and discussed how subgroup discovery tools can be configured to meet these requirements. In particular, we proposed a new quality function and a novel filtering scheme for the avoidance of redundant output. The quality function is an extension of the mean test quality that is based on the combined mean deviation of several target variables. The generally applicable filtering scheme is an incremental, i.e., higher order, quality function that is defined with respect to some desired base quality function. Its idea is to reevaluate all subgroups based on their base quality in the databases defined by their generalizations.

The quality function with several target variables is motivated by the fact that an election result is constituted by the combined results of several parties rather than just one party. Our experiments demonstrate that the introduction of several target attributes is a valuable extension: otherwise important patterns that have an interestingness resulting from the total unusualness of the results of two ore more parties can be dominated by less interesting patterns. We remark that subgroup discovery on datasets involving more than one target attribute is also known as *exceptional model mining* [13], and that our approach could thus be considered as a form of exceptional model mining (based on a new quality function).

Our other technical addition, the incremental quality function, generalizes the well-known idea of evaluating patterns with respect to their generalizations. Following our earlier specified requirements this filtering technique is completely parameter-free. This feature distinguishes our method from the other improvement-based techniques [3,7,19,20] and others, like the weighted covering scheme [12] or approaches based on affinity [5]. Note that although subgroup filtering based on the theory of relevancy [11] is also parameter-free, it only applies to data with a *binary* target variable and thus is not applicable here.

Acknowledgments

Part of this work was supported by the German Science Foundation (DFG) under the reference number ‘GA 1615/1-1’. We would like to thank the Office of City Development and Statistics for their kind cooperation.

References

1. Atzmüller, M., Puppe, F.: Semi-automatic visual subgroup mining using vikamine. J. UCS 11(11), 1752–1765 (2005)
2. Atzmüller, M., Puppe, F.: SD-map - a fast algorithm for exhaustive subgroup discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 6–17. Springer, Heidelberg (2006)
3. Bayardo, R.J., Agrawal, R., Gunopulos, D.: Constraint-based rule mining in large, dense databases. Data Min. Knowl. Discov. 4(2/3), 217–240 (2000)
4. Boley, M., Grosskreutz, H.: Non-redundant subgroup discovery using a closure system. In: ECML/PKDD, vol. (1), pp. 179–194 (2009)
5. Gebhardt, F.: Choosing among competing generalizations. Knowledge Acquisition 3, 361–380 (1991)
6. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight optimistic estimates for fast subgroup discovery. In: ECML/PKDD, vol. (1), pp. 440–456 (2008)
7. Huang, S., Webb, G.I.: Discarding insignificant rules during impact rule discovery in large, dense databases. In: SDM (2005)
8. Johnston, R., Pattie, C.: Putting Voters in Their Place: Geography and Elections in Great Britain. Oxford Univ. Press, Oxford (2006)
9. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)
10. Kralj, P., Lavrač, N., Zupan, B.: Subgroup visualization. In: Proc. 8th Int. Multiconf. Information Society, pp. 228–231 (2005)
11. Lavrac, N., Gamberger, D.: Relevancy in constraint-based subgroup discovery. In: Boulicaut, J.-F., De Raedt, L., Mannila, H. (eds.) Constraint-Based Mining and Inductive Databases. LNCS (LNAI), vol. 3848, pp. 243–266. Springer, Heidelberg (2006)
12. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. J. Mach. Learn. Res. 5(February), 153–188 (2004)
13. Leman, D., Feelders, A., Knobbe, A.: Exceptional model mining. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 1–16. Springer, Heidelberg (2008)
14. Mochmann, I.C.: Lifestyles, social milieus and voting behaviour in Germany: A comparative analysis of the developments in eastern and western Germany. PhD thesis, Justus-Liebig-University Giessen (2002)
15. Morik, K., Boulicaut, J.-F., Siebes, A. (eds.): Local Pattern Detection. LNCS (LNAI), vol. 3539. Springer, Heidelberg (2005)
16. Nijssen, S., Guns, T., Raedt, L.D.: Correlated itemset mining in roc space: a constraint programming approach. In: KDD, pp. 647–656 (2009)
17. Novak, P.K., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. J. Mach. Learn. Res. 10, 377–403 (2009)
18. Robinson, W.S.: Ecological correlations and the behavior of individuals. Am. Sociolog. Rev. (1950)
19. Webb, G., Zhang, S.: Removing trivial associations in association rule discovery. In: ICAIS (2002)
20. Webb, G.I.: Discovering significant patterns. Mach. Learn. 71(1), 131 (2008)
21. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: PKDD 1997, pp. 78–87. Springer, Heidelberg (1997)

A Description of the Data

The data used in this paper consists of 800 records, one for every polling district in the City of Cologne. Beside the number of votes obtained by the different parties in the 2009 and 2005 Bundestag elections, the records include more than 80 descriptive variables, which were assembled from different sources. The primary data source is the official city statistics, gathered and published by the Office of Statistics. Second, commercial data was used to obtain information about the type of buildings and the debt-ratio. Finally, information about the average income and the education level was taken from an anonymous citizen survey conducted by the Office of City Development and Statistics in 2008/09. The survey data is a random sample, stratified according to age, sex and urban district, which includes about 11200 responses. All variables occurring in at least one of the subgroup reported in this paper are listed in the following table. Beside the description of the variable, we also indicate the data source (OS - Official Statistics, CO - Commercial, SU - Survey).

Variable	Description
`aged_16-24`, `aged_25-34`, `aged_35-64`, `aged_65+`	age structure, i.e. the number of inhabitants aged 16-24, 25-45, etc. (OS)
`avg_living_space`	average living space per accommodation (CO)
`catholic`, `muslim`, `protestant`	number of persons with a particular religious denomination (OS)

`detached_houses`
...
`16-19_fam._buildings`
`20+_fam._buildings`
	type of buildings: number of detached houses, number of apartment buildings of different size (CO)

`education:elem._school`
`education:secondary`
`education:university`
	highest level of general education (SU)

`families`	number of families with children (OS)
`gram_school_students`	number of grammar school students (OS)

`income < 1500€`
`income 1500-3000€`
`income 3000-4500€`
`income > 4500€`
	household net income per month (SU)

`men`, `women`	percentage of male resp. female inhabitants (OS)

`occupancy 5-10 y.`
`occupancy 10-15 y.`
`occupancy 15-20 y.`
...
	duration of living in Cologne (SU)

`single_parents`	number of single parent households (OS)

`single-person_<30y`
`single-person_30-60y`
	number of one-person householders aged under 30, resp. aged 30-60 (OS)

`social_sec_claimants`	number of social security claimants (OS)