Subgroup Discovery

Martin Atzmueller

The discovery of (interesting) subgroups has a high practical relevance in all domains of science or business. For example, consider statements such as: "the unemployment rate is above average for young men with a low educational level", "smokers with a positive family history are at a significantly higher risk for coronary heart disease", or "single males living in rural areas do rarely take out a life policy". Subgroup discovery is well suited for finding such dependencies, i.e., discovering relations between a dependent and (several) independent variables, for inductive and explorative data analysis tasks. This article aims to give a first idea of subgroup discovery: we introduce the discovery task and discuss the setting and the issues concerning the search process. Then, we give an outlook on further research directions.

1 Introduction

Subgroup discovery [1, 2] is a method to identify relations between a dependent variable (target variable) and usually many explaining, independent variables. For example, consider the subgroup described by "*smoker=true* AND *family history=positive*" for the target variable *coronary heart disease=true*. Subgroup discovery does not necessarily focus on finding complete relations; instead partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient. The discovered subgroup patterns must essentially satisfy two conditions. First, they have to be interpretable for the analyst, and second they need to be interesting according to the criteria of the user. Interestingness is typically defined by a quality function, which can take certain statistical or other user-defined quality criteria into account. For example, important parameters are the difference in the distribution of a (binary) target variable concerning the subgroup and the general population, and the subgroup size.

The deviations of a subgroup from the performance of the general population are usually not simply due to statistical fluctuations, but are caused by local factors. Identifying these factors helps to understand the data in general and thus can provide useful insights for the analyst. Therefore, the main application areas of subgroup discovery are exploration and descriptive induction.

2 The Subgroup Discovery Task

A subgroup discovery task mainly relies on the following four properties: the target variable, the subgroup description language, the quality function, and the search strategy.

The target variable (e.g., coronary heart disease) may be binary, nominal or numeric. Depending on its type, there are different analytic questions, e.g., we can search for significant deviations of the mean of a numeric target variable. The description language specifies the individuals from the general population belonging to the subgroup. Subgroup description languages can be either single-relational or multi-relational. In the case of single-relational propositional languages a subgroup description can be defined as follows: Let Ω_A the set of all attributes with an associated domain $dom(a)$ of values. $\mathcal{V}_A$ is defined as the (universal) set of attribute values of the form $(a = v)$, $a \in \Omega_A, v \in dom(a)$.

Definition 1 (Subgroup Description)**.** *A subgroup description $sd = \{e_i\}$ is defined by the conjunction of a set of selection expressions. These selectors $e_i = (a_i, V_i)$ are selections on domains of attributes, $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. Ω_{sd} denotes the set of all possible subgroup descriptions.*

In our example, the subgroup "smokers with a positive family history are at a significantly higher risk for coronary heart disease" is described by the selectors *smoker=true* and *family history=positive*. The multi-relational case extends the single-relational one by including further information about the applied relational tables, using specified links between these.

A quality function measures the interestingness of the subgroup mainly based on a statistical evaluation function, e.g., the chi-squared statistical test. It is used by the search method to rank the discovered subgroups during search. Several quality functions were proposed, e.g., [2].

Definition 2 (Quality Function)**.** *A quality function $q : \Omega_{sd} \times \mathcal{V}_A \rightarrow R$ evaluates a subgroup description $sd \in \Omega_{sd}$ given a target variable $t \in \mathcal{V}_A$.*

An exemplary quality function for binary target variables is given by

$$q_{BT} = \frac{p - p_0}{\sqrt{p_0 \cdot (1 - p_0)}} \sqrt{n} \sqrt{\frac{N}{N - n}},$$

where p is the relative frequency of the target variable in the subgroup, p_0 is the relative frequency of the target variable in the total population, N is the size of the total population, and n denotes the size of the subgroup.

Then, quality functions can be used to measure the characteristics of the subgroups according to the analytical questions. In the simplest case, one population share is considered, but also several shares (segments) can be compared, e.g., segmenting by *sex = male* vs. *sex = female*.

Preprint of: Martin Atzmueller. Subgroup Discovery. Künstliche Intelligenz, 4:52–53, 2005

For the subgroup search strategy an efficient search is necessary, since the search space is exponential concerning all the possible selectors of a subgroup description.

So, a brute-force exhaustive search strategy is often not acceptable, if the search space cannot be sufficiently constrained. Therefore, non-exhaustive (heuristic) search strategies provide a better tradeoff, but often cannot guarantee to find the best solution. For subgroup discovery commonly a beam search strategy is used because of its efficiency [2]. Beam search only expands the current subgroup hypothesis further, if the specialization yields a better subgroup. E.g., Lavrac et al. [3] describe the application of the well-known propositional rule-learner *CN2* for subgroup discovery with several adaptations for the task.

Pruning the search space is a central issue for efficient subgroup discovery. Pruning can be accomplished, e.g., by utilizing the support of a subgroup, or by using special features of the quality function; for some quality functions it is possible to derive optimistic estimate functions that are used to estimate the best future quality of specializations of the current subgroup [1]. The *Apriori* algorithm for association rule mining has also been adapted for the subgroup discovery task [4] using a support threshold for pruning.

3 Further Issues

The general goal of subgroup discovery is to identify a set of high-quality subgroups. Therefore, after the brute-force phase of discovering individual subgroups, the set of all discovered subgroups needs to be refined. Criteria for selecting a set of subgroups include aiming for a small number of subgroups with a low degree of overlap of the subgroups descriptions and a high covering and quality of the combined set of subgroups. One option for subgroup selection is to apply a weighted covering approach [3] that reweights instances already included in one subgroup: it modifies the instance weights incrementally in order to focus future subgroups on the not yet covered instances in a way similar to boosting. Furthermore, (constraint-based) causal approaches [2] try to build a probabilistic network with the subgroups as nodes. Using this network, subgroups can be identified that are causal for the target group and are not redundantly dependent of other subgroups.

A further issue relates to the applied quality function which should be formulated according to the specific analytical question. Therefore, it can include both objective and subjective measures of interest. An example for a subjective measure is actionability, e.g., for decision support [5]: actionable subgroup patterns allow the decision maker to directly perform an appropriate action.

Applying visualization techniques for subgroup mining can help to make the search process more transparent in a user-guided approach [6]: both the direct search for subgroup patterns, and the selection of the final set of subgroups can be supported using appropriate visualization techniques. The quality of the discovered patterns is another important issue that can be increased using background knowledge for subgroup discovery. Background knowledge can be applied in order to reduce the number of uninteresting patterns that are discovered and also to constrain the search space, thus increasing the efficiency of the search method, e.g., [7].

4 Conclusion

This article aimed to give an overview on subgroup discovery as a quite general and pragmatic exploration approach. We introduced and discussed the main concepts for subgroup discovery and mentioned some further research issues.

As discussed in the recent literature [2, 5] subgroup discovery or subgroup mining can be applied in order to answer various analytical questions. To conclude, subgroup discovery or subgroup mining has the potential for a wide variety of applications and poses interesting and challenging issues for future research.

Kontakt

Martin Atzmueller
Künstliche Intelligenz und Angewandte Informatik
Institut für Informatik, Universität Würzburg
Am Hubland, 97074 Würzburg
Phone: +49 (0)931 888-6739, atzmueller@informatik.uni-wuerzburg.de

References

[1] Stefan Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposion on Principles of Data Mining and Knowledge Discovery*, pages 78–87, Berlin, 1997. Springer.

[2] Willi Klösgen. *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3: Subgroup Discovery. Oxford University Press, New York, 2002.

[3] Nada Lavrac, Branko Kavsek, Peter Flach, and Ljupco Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[4] Branko Kavsek, Nada Lavrac, and Viktor Jovanoski. APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. In *Proc. 5th International Symposium on Intelligent Data Analysis*, pages 230–241. Springer, 2003.

[5] Nada Lavrac, Bojan Cestnik, Dragan Gamberger, and Peter Flach. Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned. *Machine Learning*, 57(1-2):115–143, October 2004.

[6] Martin Atzmueller, Joachim Baumeister, Achim Hemsing, Ernst-Jürgen Richter, and Frank Puppe. Subgroup Mining for Interactive Knowledge Refinement. In *Proc. 10th Conference on Artificial Intelligence in Medicine*, LNAI 3581, pages 453–462, Berlin, 2005. Springer.

[7] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 647–652, 2005.