Combining Subgroup Discovery and Clustering to Identify Diverse Subpopulations in Cohort Study Data

Uli Niemann, Myra Spiliopoulou, Bernhard Preim, Till Ittermann, and Henry Völzke

Combining Subgroup Discovery and Clustering to Identify Diverse Subpopulations in Cohort Study Data

Uli Niemann, Myra Spiliopoulou, Bernhard Preim
Otto-von-Guericke University Magdeburg, Germany
Email: {uli.niemann, bernhard}@isg.cs.uni-magdeburg.de
myra@iti.cs.uni-magdeburg.de

Till Ittermann, Henry Völzke
University Medicine Greifswald, Germany
Email: {ittermann, voelzke}@uni-greifswald.de

***Abstract*—Subgroup discovery (SD) exploits its full value in applications where the goal is to generate understandable models. Epidemiologists search for statistically significant relationships between risk factors and outcome in large and heterogeneous datasets encompassing information about the participants' health status gathered from questionnaires, medical examinations and image acquisition. SD algorithms can help epidemiologists by automatically detecting such relationships presented as comprehensible rules, aiming to ultimately improve prevention, diagnosis and treatment of diseases. However, SD algorithms often produce large and overlapping rule sets requiring the expert to conduct a manual post-filtering step that is time-consuming and tedious. In this work, we propose a clustering-based algorithm that hierarchically reorganizes rule sets and summarizes all important concepts while maintaining diversity between the rule clusters. For each cluster, a representative rule is selected and then displayed to the expert who in turn can drill-down to other cluster members. We evaluate our algorithm on two cohort study datasets where the diseases hepatic steatosis and goiter serve as target variable, respectively. We report on our findings with respect to effectiveness of our algorithm and we present selected subpopulations.**

***Keywords*-classification rules clustering; subgroup discovery; medical data mining; cohort study data**

I. Introduction

Subgroup discovery (SD) aims to elucidate interesting relationships between different instances in a dataset with respect to a target variable in the form of rules [1]. In comparison with stronger, but predominantly opaque black-box models such as neural networks, support vector machines or random forests, SD algorithms trade off higher confidence with a superior interpretability of the discovered subpopulations which makes them highly applicable to domain expert-involved knowledge discovery. In epidemiological research, interesting subpopulations could be subsequently used to formulate and validate a small set of hypotheses or just to explore associations between risk factors for a specific outcome [2]. An interesting subpopulation could be phrased as "In the sample of this study, the prevalence of goiter is 32 %, while the likelihood in the subpopulation described by *thyroid-stimulating hormone* $\leq$ *1.63* mU/l $\wedge$ *body mass index* $>$ *32.5* kg/m^2 is 49%".

Manual inspection of large rule lists is tedious. A common observation is that there are groups of rules covering nearly the same set of instances. Consider for examples the antecedents of two rules: *BMI* $>$ *30* kg/m^2 $\wedge$ *waist-to-hip ratio* $>$ *1.2* and *BMI* $>$ *30* kg/m^2 $\wedge$ *waist circumference* $>$ *110* cm. These rules share probably a high number of instances based on the high correlation of anthropometric measurements and there might be multiple other similar rules describing the same *concept*. Instead of presenting *all* rules found by an SD algorithm at once, we propose to organize sets of rules in a hierarchical way such that the domain expert is able to explore a compact set of different concepts, equipped with mechanisms to drill-down to specific rules of interest.

In this study, we propose SD-Clu, a novel algorithm that combines subgroup discovery with clustering to return a set of k representative classification rules. Building up on a set of potentially highly overlapping rules generated by a SD algorithm, we leverage hierarchical agglomerative clustering to find groups of rules that cover different sets of instances. For each cluster, we nominate one rule as the group's representative that exhibits best trade-off between rule confidence and coverage towards the target variable. We evaluate our algorithm on two samples from the *Study of Health in Pomerania* investigating the diseases hepatic steatosis and goiter.

The paper is organized as follows. Section II contains related work on classification rule induction. Section III briefly reviews fundamental notations and quality measures of SD. Section IV describes the SHIP samples and introduces SD-Clu . We present our experimental results in section V. In the last section, we conclude with a summary and outlook.

II. Related Work

Multiple medical studies report on the successful application of SD algorithms where for a specific outcome value descriptive knowledge had to be derived. Examples include the extraction of potential drug targets for the treatment of dementia [3], the identification of auditory–perceptual, speech acoustic, and articulatory kinematic characteristics that are predictive for subpopulations of preschool children

with speech sound disorders [4], and the discovery of characteristics in patient subpopulations with different psychiatric emergency department admission times [5].

However, SD techniques often return a set of overlapping rules which are too large to be manually inspected by the domain expert. Therefore, it is necessary to reduce or filter the set of rules to retain the most representative ones. Popular SD algorithms such as SubgroupMiner [6], SD [7] and CN2-SD [8] define a fixed beam width to limit the number of further expanded subgroup candidates at each iteration. Often, a post-pruning step is applied to reduce the cardinality of the rule set to a fixed k using a quality criterion such as weighted relative accuracy or significance. Even if both beam width search and top-k pruning are applied, the end result often still contains redundant rules. This is due to dependencies between the (non-target) variables, which lead to large numbers of variations of a particular finding. Especially top-k pruning leads to different variations of the same concepts (in other words large instance overlap) [9].

In contrast to SD, predictive rule learning algorithms such as CN2 [10] or RIPPER [11] iteratively generate a rule, then remove covered instances from the training set and learn another rule that again covers some of the remaining instances until no instances remain, following a separate-and-conquer strategy [12]. These algorithms often induce a decision list where each rule covers a unique set of instances. This "exclusiveness" property leads to the problem that important concepts may remain uncovered. For example, the algorithm might induce a rule involving instances with a high BMI but might not be able to find a slightly weaker association with income since these instances are immediately removed from the instance candidate space. Furthermore, the rules produced in each iteration have increasingly lower support. This is problematic in epidemiological studies: rules with low support are not actionable and may represent artifacts of the study sample.

Instead of completely removing the instances covered by each rule, before building the next rule, weighted covering approaches utilize information on how often instances have been covered to influence the selection of the next antecedent (e.g. [8]). Even with weighted covering, important concepts might not be captured. Also, weighted covering introduces a new parameter controlling weight decrease which is difficult to set especially for domain experts. Further, both traditional predictive rule learning algorithms and their weighted covering extensions introduce order dependencies among rules. A rule is dependent on all of its previous rules and the instances of the target variable they cover, and it may not be meaningful or significant when interpreted individually.

III. Fundamentals

A. Subgroup Discovery

Our method builds on subgroup discovery (SD) and agglomerative hierarchical clustering. SD algorithms induces descriptions of subgroups of the data (subpopulations) that exhibit an "interesting" behavior where interestingness is formalized by a quality function. Subgroup discovery algorithms induce rules in the form of $r : Antecedent \rightarrow T = v$, (e.g. $(TSH \leq 1.63 \wedge BMI > 32.5) \rightarrow Goiter = POS$) where v is the requested value for the target variable T and $Antecedent$ is a conjunction of variable-value pairs able to describe a statistical distribution with respect to T that considerably deviates from the total population. In this paper, we study SD on a binary target variable problem (positive or negative outcome). A subpopulation of r, $s(r)$ is the set of instances that satisfy the antecedent of a rule r, also known as cover set of r.

B. Hierarchical Clustering

Agglomerative Hierarchical Clustering iteratively merges similar instances to clusters, in a bottom-up way. The order of merging two clusters depends on the linkage strategy: in "complete linkage" the distance between two clusters is defined as the distance between two instances (one per cluster) that are most far apart from each other. The two clusters that minimize this maximum distance are selected for merging. Using a dendrogram it is possible to drill down the "tree" of clusters to understand the progressive merging process. Optionally, a parameter k can be specified to obtain a specific partitioning.

IV. Materials & Methods

A. Definitions

Let $r : Antecedent \rightarrow Consequent$ be a classification rule. We define as "subpopulation of r", $s(r)$, the set of instances that satisfy the rule's antecedent, also known as cover set of r. We denote as $|s(r)_{T=v}|$ the number of instances that additionally exhibit the target variable value of interest, where T stands for *target* variable. The coverage of r, which is the fraction of instances covered by r is then defined as $Cov(r) = |s(r)|/n$. The support of r quantifies the percentage of instances covered by r with $T = v$, calculated as $Sup(r) = |s(r)_{T=v}|/n$. The confidence of r (also referred to as precision or accuracy) is defined as $Conf(r) = |s(r)_{T=v}|/|s(r)|$ and measures the relative frequency of instances satisfying the complete rule among those satisfying only the antecedent. The Weighted Relative Accuracy of a rule balances coverage and confidence gain and is often used as internal quality criterion for candidate generation [1]. It is defined as $WRAcc(r) = Cov(r) \cdot (Conf(r) - \frac{n_{T=v}}{n_{T \neq v}})$.

B. Cohort Study Data

We consider two multi-factorial diseases, hepatic steatosis and goiter, using data from the "Study of Health in Pomerania" (SHIP) [13]. SHIP encompasses two independent cohorts consisting of participants aged from 20 to 79 years with main residency in the study region. Baseline examinations

for the first cohort were performed between 1997 and 2001 (SHIP-0, n=4308). Three follow-up examination were done in intervals of 5-years, continuously adding new variables including MRI scans beginning from the second follow-up (SHIP-2). Baseline information for the second, independent cohort (TREND-0, n=4420) was collected in 2008-2012.

The samples `HepStea` and `Goiter` contain for each participant variables from a personalized interview, on medication intakes, blood and urine laboratory testings, whole-body magnet resonance imaging (MRI), anthropometric, bioelectrical impedance analyses, echocardiography, carotic ultrasound, liver ultrasound, ECG, blood pressure measurements, breath gas analyses, bone density measurements, grip strength testings and spiroergometry.

For `HepStea`, we derive the binary target variable by discretizing the liver fat accumulation obtained from the MRI report. Study participants with a liver fat accumulation not exceeding 10 % are mapped to the negative class `NEG`, values greater than 10 % are mapped to the positive class `POS` to denote absence or presence of the disease. Out of the random sample of 886 participants for which the MRI report of SHIP-2 is available, 694 belong to `NEG` (78.3 %) and 192 to `POS` (21.7 %). We consider 99 variables, chosen exclusively from SHIP-0, so that we can assess their long-term impact, as expressed 10 years later in SHIP-2.

The target variable for `Goiter` was derived by thyroid volume assessment via ultrasonography. Goiter was defined for a thyroid volume exceeding 18 mL in women and 25 mL in men [14]. Out of the 4400 participants for which the target variable is available in TREND-0, 3010 belong to `NEG` (68.4 %) and 1390 to `POS` (31.6 %). Apart from the target variable, we use a total of 182 variables that were pre-selected by a medical expert as potential risk factors.

C. SD-CLU

We propose a clustering algorithm, SD-CLU, which organizes the discovered rules into a dendrogram, extracts a set of clusters and maps each cluster into a *proxy* rule. For the clustering of rules, we define rule similarity on the basis of the mutually covered instances: To quantify this similarity, we adapt the DICE coefficient [15], so that for two rules r_1, r_2 with corresponding subpopulations $s(r_1), s(r_2)$ we compute

$$d(r_1, r_2) = 1 - \frac{2\,|s(r_1) \cap s(r_2)|}{|s(r_1)| + |s(r_2)|} \tag{1}$$

as dissimilarity function (two rules are identical when their dissimilarity is zero). The set of clusters to be extracted from the dendrogram can be defined as parameter k to be input by the user. Alternatively, this number can be derived with help of a cluster quality function. To this purpose, we define the silhouette coefficient for a set of clusters ξ over a set of rules R as

$$Silh(R, \xi) = \frac{1}{|R|} \sum_{r \in R} \frac{b(r) - a(r)}{\max\{a(r), b(r)\}}, \tag{2}$$

where $a(r) = \frac{\sum_{y \in Y} d(r,y)}{|Y|-1}$ is the average dissimilarity between r and the other rules in the cluster $Y \in \xi$ which contains r, while $b(r) = \frac{\sum_{y \in Z} d(r,y)}{|Z|}$ is the average dissimilarity between r and the rules in the cluster $Z \in \xi$ which is the closest to the cluster Y containing r. Then, we traverse the dendrogram bottom-up, compute the silhouette for each set of clusters ξ and select as ξ_{opt} the set of clusters with the best silhouette value. The optimal number of clusters is then the cardinality $|\xi_{opt}|$.

Finally, we map each cluster $Y \in \xi_{opt}$ to a representative rule. To do so, we invoke WRAcc (cf. Sec. III) for each rule $r \in Y$ and select as cluster proxy $cp(Y)$ the rule for which WRAcc is maximum.

D. Representativeness of a set of cluster proxies

Let R be a set of rules and let ζ be a set of clusters over R. Typically, ζ will be the optimal set of clusters, computed as described in the previous subsection, but it can be any set of clusters chosen by the user, as long as it contains all rules in R. For ζ, let $R_\zeta = \{cp(Y) | Y \in \zeta\}$ denote the set of cluster proxy rules, according to the definition of cluster proxy rule $cp()$ in the previous subsection. To quantify how representative such a set of rules is, we proceed as follows. First, let $U \subseteq R$ be an arbitrary subset of the complete set of rules, and let x be an instance. We define the *coverage rate* of x towards U as

$$covRate(x, U) = \frac{\sum_{r \in U} isCovered(x, r)}{|U|} \tag{3}$$

where $isCovered(x, r)$ is equal to 1, if x covers r, and 0 otherwise. We observe that for a set of rules U, an instance x cannot be covered by more than $|U|$ rules. Let R_x be the set of rules that cover instance x, i.e. for $R_x = \{r \in U | isCovered(x, r) = 1\}$. For the whole set of instances X we now create bins: $bin_i(U) = \{x \in X | |R_x| = i\}$. Further, let $bin_0(U) = \{x \in X | \forall r \in U : isCovered(x, r) = 0\}$. Evidently, an instance x can be covered by $0, 1, \ldots, |U|$ rules, i.e. $covRate(x, U)$ can take one of $|U| + 1$ values. In contrast, $covRate(x, R)$ can take one of $|R| + 1$ values, a usually much larger number. We therefore map the possible values of $covRate(x, R)$ into the much smaller set of possible values by rounding, computing

$$adjCovRate(x, U, R) = \frac{Round(covRate(x, R) \cdot |U|)}{|U|}. \tag{4}$$

Then, for the complete set of instances X, a set of induced rules R, the clustering ζ over R and the set of cluster proxy rules R_ζ, the *representativeness* of R_ζ is defined as

$$representativeness(R_\zeta, R) = 1 - \frac{\sum_{x \in X} |adjCovRate(x, U, R) - covRate(x, R_\zeta)|}{|X|}. \tag{5}$$

V. EXPERIMENTS

A. SD Algorithms

For subgroup discovery we employ the algorithms HotSpot [16] and SD-Map [17]. HotSpot is a beam width SD algorithm that implements a level-wise top-down approach for extracting rules. At each iteration, only the b highest ranked candidates are generated according to confidence. We use the implementation from the Waikato Environment for Knowledge Analyis (WEKA). SD-Map is an exhaustive SD algorithm that adapts the popular FP-growth association rule learning method [18]. It prunes rules that fall below a minimum coverage threshold. It implements a depth-first search for candidate generation and ranks induced rules according to a given quality function. We use the implementation from the VIKAMINE framework [19].

B. Parameter Settings

The implementation of SD-Map can handle only categorical variables. Therefore, we convert all numeric variables using the MDL-based (supervised) discretization of Fayyad et al. [20]. For SD-Map, we set the minimum coverage threshold to 0.05 to avoid too small, overfitting rules. We leverage WRAcc as quality function and define a minimum threshold value of 0.025. For HotSpot, we set the support threshold of a rule to exceed 0.05. The beam width is set to 500. Further, to avoid rather meaningless literals, we restrict expanding of a rule body with another literal to a relative confidence gain of at least 0.3.

C. Results

Fig. 1 shows the optimal number of clusters for each dataset and SD algorithm. Table I depicts the optimal k and the fraction of rules shown to the analyst. For example, the clustering with optimal silhouette coefficient for the algorithm HotSpot on the dataset `Goiter` has 76 clusters and thus 76 rule cluster proxies (cf. Table I, 3rd row, 5th column), which makes only 21.3 % of the total number of rules (cf. 5th row, last column). Thus, by showing the expert only the RCP in the beginning, the amount of time (s)he will spend inspecting the rules will decrease.

The optimal cardinality of ζ_{opt} shown in Fig. 1 could be used as a suggestion, but the expert is free to specify the number of rules (s)he wants to obtain. For instance, if the expert feels $|\zeta_{opt}| = 100$ for HotSpot on `HepStea` is too large, the diagram would show him that a reduction to $|\zeta_{opt}| = 58$ is possible where $Silh$ reduces just slightly from 0.48 to 0.37. Also in the other direction: if $|\zeta_{opt}|$ is rather low, the diagram indicates that a minor increase does not change $Silh$ strongly; thus the added rules may be important also. The expert could even analyze the diagram to derive a range instead of a single value, e.g. the range where $Silh$ is above 90 % of its maximum.

To assess representativeness of rule cluster proxies, we compare them to the output of three baseline algorithms.

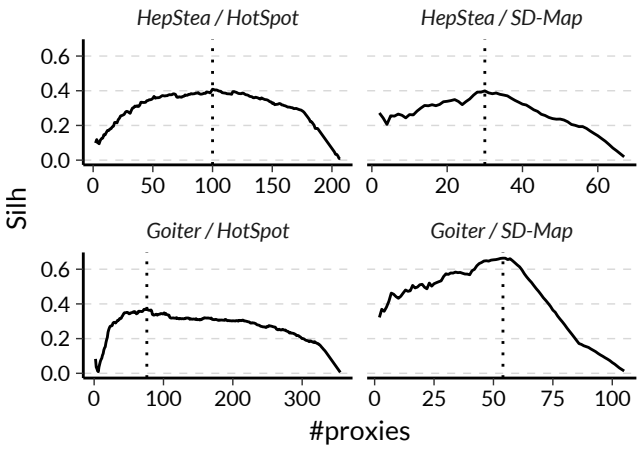

Figure 1. Silhouette coefficients ($Silh$) of SD-CLU using complete linkage for each dataset/ algorithm combination. The k of the clustering with the highest $Silh$ score is indicated by a dotted vertical line.

Table I
BEST $Silh$ FOR EACH DATASET-ALGORITHM COMBINATION.

Dataset	Algorithm	$\|R\|$	$Silh(\zeta_{opt})$	$\|\zeta\|$	$\frac{\|\zeta\|}{\|R\|}(\%)$
`HepStea`	HotSpot	208	0.41	100	48.1
`HepStea`	SD-Map	68	0.40	30	44.1
`Goiter`	HotSpot	356	0.37	76	21.3
`Goiter`	SD-Map	106	0.66	54	51.6

In particular, we vary the number of rules k returned to the expert as follows: we sort the rules in R on Odds Ratio (baseline 1), on Coverage (Baseline 2), on WRAcc (Baseline 3). In Fig. 2, we depict the three baselines (bottom to top) and our approach, varying the number of rules k returned to the expert for the HotSpot algorithm on the `HepStea` dataset. The horizontal axis of each diagram shows the number of representative rules k returned to the expert. The vertical axis of each diagram shows the $adjCovRate$ of the instances w.r.t. R (solid black curve) and with respect to the subset or representative rules (points); the deviation between the two is represented by the dark gray area between the solid black curve and the dotted curve which is a locally weighted scatterplot smoothing regression of the points.

Fig. 2 illustrates $representativeness$ of ζ_{opt} for `HepStea` / HotSpot and juxtaposes it to the baseline approaches. For all approaches, $representativeness$ improves with increasing number of representative rules k. For example, $representativeness$ increases from 0.78 to 0.98 for SD-CLU and $k = 5$ to $k = 100$ which means that the absolute difference between $adjCovRate$ of ζ and $covRate$ of R over all instances successively decreases. However, for a given k, the baseline approaches' sets of representative rules are less representative than SD-Clu's proxy rules, e.g. 0.91, 0.92, 0.91 vs. 0.96 for $k = 50$, respectively (cf. 3rd column of plot matrix in Fig. 2).

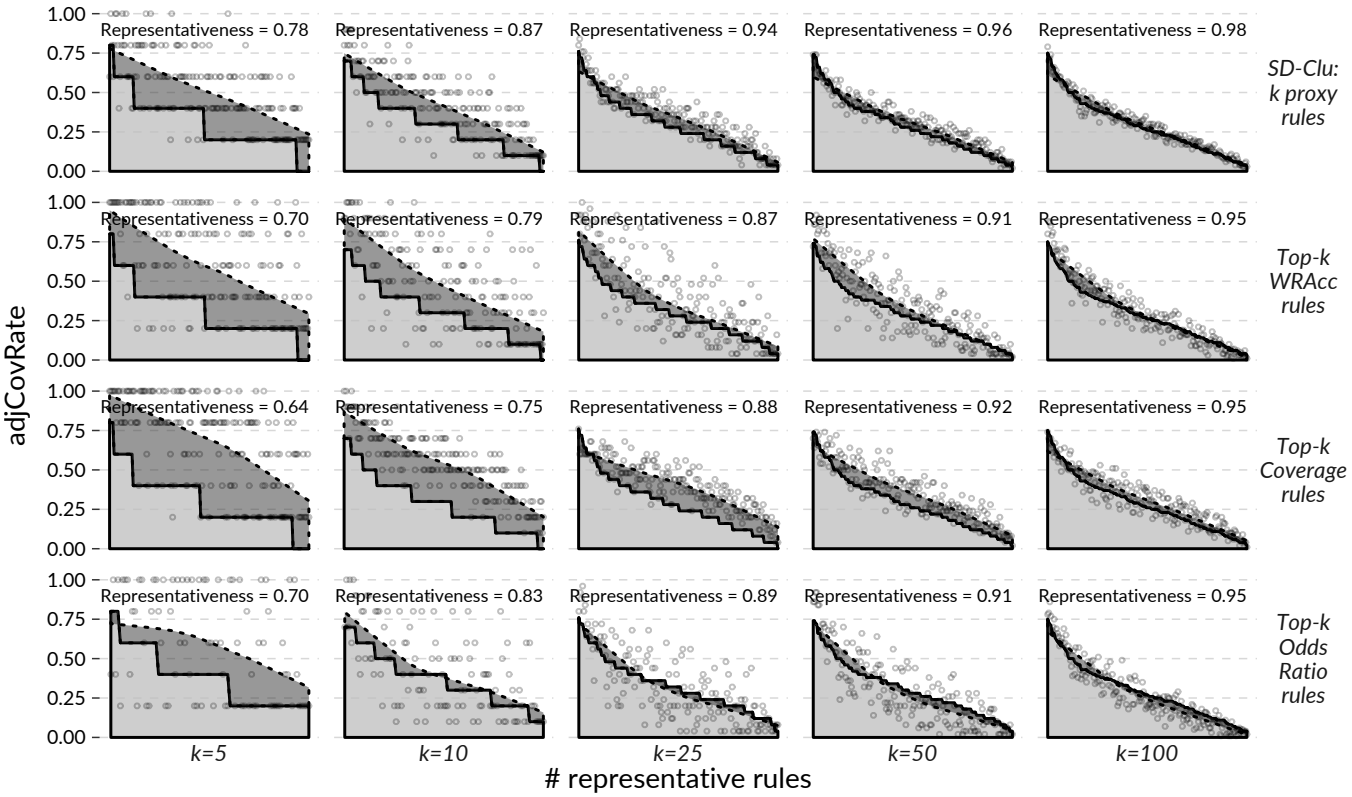

Figure 2. Comparison of $representativeness$ between SD-CLU and three baseline approaches for different k (`HepStea` / HotSpot). Instances are sorted by $adjCovRate$ w.r.t. R in descending order, shown by the light gray filled area below the solid black curve. Points depict $adjCovRate$ of an instance regarding the set of representative rules of the approach (row) and k (column). A LOWESS regression is learned on the points and shown as dotted curve. The deviation between solid and dotted curve indicates the representativeness of the top-k representative rules of the approach. This is illustrated by the dark gray filled area in-between solid curve and dotted curve where smaller areas are better and reflect higher $representativeness$ values.

D. Findings

Table II depicts the subpopulations of 5 rule cluster proxies, selected from the proxies found by one of the two algorithms for `HepStea` (cf. Table I, 5th column, 1-2th row), while Table III shows a choice of 5 rule cluster proxies from the ones found by both algorithms for `Goiter` (cf. Table I, 5th column, 3-4th row). The prevalence of hepatic steatosis, resp. goiter, in each of these subpopulations is significantly higher than in the corresponding complete population. These subpopulations are characterized by known risk factors for hepatic steatosis (cf. 2nd column in Table II): large waist circumference and BMI, high alcohol consumption, lack of sleep, genetic predisposition, and high scores in some of the medical tests (ALAT and LDL). Similarly, for goiter (cf. 2nd column in Table III), the identified subpopulations are characterized by medical test scores that are indicative of low TSH.

VI. CONCLUSION

In this paper, we have tackled the problem of high instance-overlap in sets of rules generated by subgroup discovery algorithms. We presented SD-CLU, an algorithm that hierarchically clusters rules to yield groups of rules that cover different sets of instances. For each cluster, a representative rule is nominated, limiting the number of rules displayed to the domain expert and thus reducing the amount of time spent for rule inspection. Further, we have introduced a *representativeness* measure that assesses whether instances are similarly often covered by representatives as by the total rule set. We have evaluated SD-CLU on two samples from an epidemiological study where we have extracted an optimal set of *proxy* rules (i) that contains considerably less rules than the total rule set and (ii) is more representative in comparison with the baseline approaches, respectively. In future work, we would like to embed SD-CLU into an application to conduct a comprehensive domain-expert evaluation w.r.t usability and applicability of the algorithm.

REFERENCES

[1] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and information systems*, vol. 29, no. 3, pp. 495–525, 2011.

[2] U. Niemann, H. Völzke, J.-P. Kühn, and M. Spiliopoulou, "Learning and inspecting classification rules from longitu-

Table II
SELECTED PROXY RULES FOR HEPSTEA LEARNED ON THE POS. OUTCOME (BOTH ALGORITHMS). OR: ODDS RATIO, P-VALUE: χ^2 TEST P-VALUE.

#	Rule Cluster Proxy Antecedent	Cov ▽	OR	WRAcc	p-value
1	waist circumference = '$>$ 88.15 cm' $\wedge$uric acid= '$>$ 278.5 μmol/l $\wedge$ LDL = '$>$ 2.46 mmol/l'	0.28	2.9	0.05	1.7e^{-172}
2	BMI = '$>$26.3 kg/m^2' $\wedge$ ALAT = '$>$ 0.385 μkatal/l' $\wedge$ diastolic BP = '$>$ 79.7 5 mmHg'	0.25	3.3	0.05	1.4e^{-184}
3	SF-12 physical health score $\leq$ 47.3 $\wedge$ ALAT = $>$ 0.39 $\mu katal/l$	0.12	2.8	0.03	4.0e^{-241}
4	intima-media thickness $\leq$ 0.79 mm $\wedge$ left ventricular mass index $>$ 40.74 g/m $\wedge$ uric acid $>$ 262 μmol/l	0.10	4.0	0.03	8.5e^{-252}
5	sleep problems = TRUE $\wedge$ alcohol problems = TRUE $\wedge$ excessive waist circumference = TRUE	0.04	4.6	0.01	1.5e^{-289}

Table III
SELECTED PROXY RULES FOR GOITER LEARNED ON THE POS. OUTCOME (BOTH ALGORITHMS). OR: ODDS RATIO, P-VALUE: χ^2 TEST P-VALUE.

#	Rule Cluster Proxy Antecedent	Cov ▽	OR	WRAcc	p-value
1	hip circumference = '$>$98.75 cm' $\wedge$ age = '$>$38.5 years' $\wedge$ social phobia = FALSE	0.42	1.7	0.05	1.6e^{-98}
2	ECG: P duration = '$>$111 ms' $\wedge$ intima-media thickness = '$>$0.55 mm' $\wedge$ ATC_A02BC intake = FALSE	0.35	1.8	0.05	2.0e^{-119}
3	bone density: sonography speed of sound left $\leq$ 1522.6 m/sec $\wedge$ TSH $\leq$ 0.664 mU/l $\wedge$ hip circumference $\leq$ 104.2 cm	0.02	∞	0.01	5.4e^{-284}
4	TSH $\leq$ 0.251 mU/l	0.02	8.7	0.01	2.8e^{-279}
5	education level = 'Univ./College' $\wedge$ height $>$ 178 cm $\wedge$ methane $>$ 0.622 parts per billion	0.02	7.5	0.01	3.1e^{-281}

dinal epidemiological data to identify predictive features on hepatic steatosis," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5405–5415, 2014.

[3] T.-P. Nguyen, C. Priami, and L. Caberlotto, "Novel drug target identification for the treatment of dementia using multi-relational association mining," *Scientific reports*, vol. 5, 2015.

[4] J. C. Vick, T. F. Campbell, L. D. Shriberg, J. R. Green, K. Truemper, H. L. Rusiewicz, and C. A. Moore, "Data-driven subclassification of speech sound disorders in preschool children," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 6, pp. 2033–2050, 2014.

[5] C. J. Carmona, P. González, M. Del Jesus, M. Navío-Acosta, and L. Jiménez-Trevino, "Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department," *Soft Computing*, vol. 15, no. 12, pp. 2435–2448, 2011.

[6] W. Klösgen and M. May, "Spatial subgroup mining integrated in an object-relational spatial database," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2002, pp. 275–286.

[7] D. Gamberger and N. Lavrac, "Expert-guided subgroup discovery: Methodology and application," *Journal of Artificial Intelligence Research*, vol. 17, pp. 501–527, 2002.

[8] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski, "Subgroup discovery with cn2-sd," *Journal of Machine Learning Research*, vol. 5, pp. 153–188, 2004.

[9] M. van Leeuwen and A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 208–242, 2012.

[10] P. Clark and T. Niblett, "The cn2 induction algorithm," *Machine learning*, vol. 3, no. 4, pp. 261–283, 1989.

[11] W. W. Cohen, "Fast effective rule induction," in *International Conference on Machine Learning*, 1995, pp. 115–123.

[12] J. Fürnkranz, D. Gamberger, and N. Lavrač, *Foundations of rule learning*. Springer Science & Business Media, 2012.

[13] H. Völzke, D. Alte, C. O. Schmidt, D. Radke *et al.*, "Cohort profile: The Study of Health in Pomerania," *International Journal of Epidemiology*, vol. 40, pp. 294–307, 2011.

[14] R. Gutekunst, W. Becker, R. Hehrmann, T. Olbricht *et al.*, "Ultrasonic diagnosis of the thyroid gland," *Deutsche medizinische Wochenschrift*, vol. 113, no. 27, pp. 1109–1112, 1988.

[15] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD expl. newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[17] M. Atzmüller and F. Puppe, "SD-Map–A fast algorithm for exhaustive subgroup discovery," in *Proc. of PKDD*. Springer, 2006, pp. 6–17.

[18] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.

[19] M. Atzmüller and F. Lemmerich, "Vikamine–open-source subgroup discovery, pattern mining, and analytics," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 842–845.

[20] U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," in *Proceedings of the 13th International Joint Converence on Artificial Intelligence*, vol. 2, 1993, pp. 1022–1027.