On the Uniform Control of the Vapnik-Chervonenkis Dimension in Subgroup Discovery Using Formal Concept Analysis

Georg Schollmeyer[1]

[1] *Department of Statistics, LMU Munich e-mail:* georg.schollmeyer@stat.uni-muenchen.de

Very preliminary draft!

Abstract:

In this paper we analyze two methods of regularization in subgroup discovery. The first approach is to simply constrain the description length of the envisaged subgroup descriptions. The second approach is in the spirit of formal concept analysis: One firstly identifies large contranominal scales in the data structure, which correspond to large shatterable sets and thus a high VC dimension of the space of subgroups. Then, the impact of the largest contranominal scales is reduced by temporarily deleting the corresponding objects to obtain a smaller and particularly less complex subcontext of the given formal context. Based on the subcontext, we analyze two variants of reincorporating the deleted objects: The first variant looks at the closed item sets of the subcontext and their induced subgroups in the original formal context. The second variant instead restructures the original context by projecting the removed objects onto the concept lattice of the subcontext. We analyze the behavior of the methods with respect to a local version of the VC dimension and the actual size of the regularized subgroup spaces. It turns out that for the first method, the VC dimension cannot be guaranteed to be controlled, but one can still obtain generalization bounds through the analysis of the size of the regularized subgroup space. While for the first variant of the second method the situation is similar, the second variant is able to (uniformly) control the VC dimension to a given nominal dimension K. Finally, we analyze the three methods also by means of a data example from the German General Social Survey (GGSS).

Keywords and phrases: VC Dimenson, Formal Concept Analysis, Subgroup Discovery, Regularization.

1. Introduction

In subgroup discovery, one important aspect is the question if the interestingness of the discovered subgroup(s) is in some certain sense statistically sound. The purely statistical analysis of the methodology of subgroup discovery seems to be still in its infancy (cf., e.g., [7, 3]). One very simple way of regularizing the generally ill-conditioned subgroup discovery problem would be to reduce the space of envisaged subgroups by restricting the length of the subgroup descriptions to a certain maximal length K. In this contribution we analyze how

such a restriction of the description length exactly reduces the space of subgroups in terms of the Vapnik-Chervonenkis dimension (VC dimension) of the subgroup space. We describe here subgroups in the language of formal concept analysis (FCA), which particularly allows the common conceptual treatment of data with different scales of measurement (especially dichotomous, categorical and ordinal/interordinal scaled data) via conceptual scaling, as well as the analysis of dualities between data points (i.e. objects in terms of FCA) and attributes. It turns out that by restricting the description length, the space of subgroups is only reduced for that subdomains which (in a certain sense) suffer from a high VC dimension, which appears to be quite reasonable. On the other hand, it is not guaranteed that every domain of high local VC dimension will be regularized sufficiently (in the sense that the local VC dimension of the regularized space of subgroups is lower than or equal to K). However, one can still show a generalization inequality based on the size of the regularized subgroup space. Furthermore, if one does not restrict the description length but instead the number of statistical units which can generate the envisaged subgroup, then one can also show a generalization inequality, which is comparable to the generalization inequality one would obtain with a VC dimension K. Additionally, we propose another way of regularization which uses insights from formal concept analysis and firstly identifies large shatterable sets in terms of identifying large substructures which constitute contranominal scales. This way of regularizing the subgroup discovery problem also leads to a regularization of only that subdomains, which (in a certain, now slightly different sense) suffer from a high VC dimension. A further modification of the approach eventually also ensures that the VC dimension is globally (and thus of course also locally) bounded by K. The paper is organized as follows: In Sections 1.1 to 1.3 we introduce the needed basics of formal concept analysis, subgroup discovery and Vapnik-Chervonenkis theory. In Section 2 we present and analyze the two methods for the regularization of the subgroup discovery problem. Section 3 analyzes the two methods of regularization by means of a data example from the German General Social Survey (GGSS) while Section 4 concludes.

1.1. Formal concept analysis

Formal concept analysis (FCA) is an applied mathematical theory rooted in an attempt to mathematically formalize the notion of a *concept*. Concretely, in formal concept analysis one starts with a so-called formal context $\mathbb{K} = (G, M, I)$ where G is a set of objects, M is a set of attributes and $I \subseteq G \times M$ is a binary relation between the objects and the attributes with the interpretation $(g, m) \in I$ iff object g has attribute m. If $(g, m) \in I$ we also use infix notation and write gIm. In the context of statistical data analysis, the objects typically are the data points, for example the persons that participated in a survey.The attributes are the observed values of the interesting variables, for example the answer *yes* or *no* to the posed questions and gIm means that person g answered the question m with *yes*. (If the answers to the questions in a survey are not

binary, then one can transform them into binary attributes with the method of conceptual scaling.) A formal concept of the context $\mathbb{K}$ is a pair (A, B) of a set $A \subseteq G$ of objects, called extent, and a set $B \subseteq M$ of attributes, called intent, with the following properties:

i) Every object $g \in A$ has every attribute $m \in B$ (i.e.: $\forall g \in A \forall m \in B : gIm$).
ii) There is no further object $g \in G \backslash A$ that has also all attributes of B (i.e.: $\forall g \in G : (\forall m \in B : gIm) \Longrightarrow g \in A$).
iii) There is no further attribute $m \in M \backslash A$ that is also shared by all objects $g \in A$ (i.e. $\forall m \in M : (\forall g \in A : gIm) \Longrightarrow m \in B$).

The property of being a formal concept can be characterized with the operators $\Phi : 2^M \longrightarrow 2^G : \quad B \mapsto \{g \in G \mid \forall m \in B : gIm\}$ and $\Psi : 2^G \longrightarrow 2^M : A \mapsto \{m \in M \mid \forall g \in A : gIm\}$ as (A, B) is a formal concept $\iff \Psi(A) = B \,\&\, \Phi(B) = A$. In the sequel, we will abbreviate both Ψ and Φ with $'$. (Which of the two operators is meant will be always clear from the context.) The set of all formal concepts of a given context $\mathbb{K}$ is denoted with $\mathfrak{B}(\mathbb{K})$ and, equipped with the sub-concept relation $(A, B) \sqsubseteq (C, D) : \iff A \subseteq C \& B \supseteq D$ builds a complete lattice which is called the concept lattice of $\mathbb{K}$. The set of all extents of $\mathbb{K}$ is denoted with $\mathfrak{E}(\mathbb{K})$ and the set of all intents is denotetd with $\mathfrak{I}(\mathbb{K})$. A formal attribute-implication is a pair (Y, Z) of subsets of M, also denoted by $Y \longrightarrow Z$. We say that an implication $Y \longrightarrow Z$ is valid in a context $\mathbb{K}$ if every object that has all attributes in Y does also have every attribute in Z. We say that a set $B \subseteq M$ respects an implication $Y \longrightarrow Z$, if it is no superset of Y or if is a superset of Z. The set of all intents is then exactly the set of all subsets $B \subseteq M$ that respect all implications that are valid in the context. A set $B \subseteq M$ of attributes is called (attribute-)implication-free, if there are no two disjoint nonempty subsets $Y, Z \subseteq B$ for which the (attribute-)implication $Y \longrightarrow Z$ is valid in the context $\mathbb{K}$. Dual to this, a formal object-implication is a pair (Y, Z) of subsets of G, also denoted by $Y \longrightarrow Z$, and we say that an (object-) implication $Y \longrightarrow Z$ is valid in a context $\mathbb{K}$ if every attribute that is shared by all objects in Y is also shared by every object in Z. A set $A \subseteq M$ of objects is called (object-)implication-free, if there are no two disjoint nonempty subsets $Y, Z \subseteq A$ for which the (object-)implication $Y \longrightarrow Z$ is valid in the context $\mathbb{K}$.

1.2. The subgroup discovery problem in the language of formal concept analysis

The basic task of subgroup discovery can be stated as:

"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer,...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting" i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest." [8]

Translated into the language of formal concept analysis, the problem of subgroup discovery reads as: Given a formal context $\mathbb{K} = (G, M, I)$, a number

$k \in \mathbb{N}$, a target variable $t : G \longrightarrow \mathcal{A}$ with values in a set $\mathcal{A}$, and given a so-called quality function $q : 2^G \longrightarrow \mathbb{R}$, which quantifies the interestingness of subgroups w.r.t. the target variable t, the task of subgroup discovery is to find that k formal concepts (A, B), for which the value $q(A)$ of the quality function is the largest. In the following, we restrict the setting to $k = 1$ and the Piatetsky-Shapiro quality function q_{PS} for a binary target variable t (cf., [4]). For $t(g) \in \{0, 1\}$ this quality function is defined as $q_{PS}(A) = n \cdot (p - p_0)$, where $n = |A|$, p is the proportion of statistical units with target value 1 within the whole group and p_0 is the proportion of units with target value 1 within the subgroup A.

1.3. Vapnik-Chervonenkis theory

Let V be a basic set and let $\mathcal{S} \subseteq 2^V$ be a family of subsets of V. For an arbitrary set $A \subseteq V$ define the projection of $\mathcal{S}$ onto A as $\mathcal{S}_A := \{A \cap B \mid B \in \mathcal{S}\}$. The growth function $m^{\mathcal{S}}$ ist defined as $m^{\mathcal{S}} : \mathbb{N} \longrightarrow \mathbb{N} : r \mapsto \max\limits_{\substack{A \subseteq V, \\ |A| = r}} |\mathcal{S}_A|$. A set $A \subseteq V$ is called shatterable w.r.t. $\mathcal{S}$ if $\mathcal{S}_A = 2^A$. The Vapnik-Chervonenkis dimension (VC dimension) is defined as the highest cardinality of a shatterable set. Then, for a finite V, if the VC dimension of $\mathcal{S}$ is d, we have $|\mathcal{S}| \leq \sum\limits_{i=0}^{d} \binom{|V|}{i} \leq |V|^d + 1$, which implies that also for a family $\mathcal{S}$ over a possibly infinite basic set V with VC dimension d we have $m^{\mathcal{S}}(r) \leq \sum\limits_{i=0}^{d} \binom{r}{i} \leq r^d + 1$. Now, define the quantity $\Gamma(r, d) := \sum\limits_{i=0}^{d} \binom{r}{i}$. With this, we can state the sufficient condition for uniform convergence (cf., [5, 6]):
Let $(X_1, \ldots, X_n)$ be an i.i.d.-sample of random variables with values in V and marginal image law P, joint image law $P^{\otimes n}$, and an associated (marginal) empirical law ν_n. Define the statistic $D_{\mathcal{S},n} := \sup\limits_{A \in \mathcal{S}} |\nu_n(A) - P(A)|$. Given that $n \geq 2/\varepsilon^2$, which we will always assume in the sequel, we have the generalization inequalities

$$P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot m^{\mathcal{S}}(2n) e^{-\frac{n\varepsilon^2}{4}} \quad \text{and} \quad P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot \Gamma(2n, d) e^{-\frac{n\varepsilon^2}{4}}.$$

In the sequel, we will analyze the family of all extents of a given context $\mathbb{K}$, i.e.: $\mathcal{S} := \mathfrak{E}(\mathbb{K})$ for a given context $\mathbb{K}$. Furthermore, we assume that the set $G = V$ is exactly the set of the covariates of all observed statistical units in a sample and we are interested in the statistical inference about the unknown underlying population. We will analyze different methods for the regularization of $\mathcal{S}$ by making $\mathcal{S}$ smaller and/or less complex in terms of the VC dimension. In the context of formal concept analysis, where $\mathcal{S}$ is a closure system (i.e. closed under arbitrary intersections), the shatterable sets are easier to analyze compared to general families of sets: A set $A \subseteq G$ is shatterable if and only if it is object-implication-free. Additionally, a set $A \subseteq G$ is shatterable if and only if there exists a set $B \subseteq M$, such that the subcontext $(A, B, I \cap A \times$

B) builds a contranominal scale, which means that for all $a \in A$ there exists exactly one $b \in B$ with $(a, b) \notin I$ and vice versa. This particularly implies via duality that the VC dimension of the family of all extents of a context equals the VC dimension of the family of all intents of the context. Therefore, we call this VC dimension also the VC dimension of the context. The statistic $D_{\mathcal{S},n}$ in the Vapnik-Chervonenkis-setting is closely related to the Piatetsky-Shapiro quality function q_{PS} for subgroup discovery, one can easily show that $q_{PS} = C \cdot \sup_{A \in \mathcal{S}} \left(\nu^1(A) - \nu^2(A)\right)$ where $C > 0$ is a constant, ν^1 is the empirical measure w.r.t. the subpopulation of individuals with target value 1 and ν^2 is the empirical measure w.r.t. the subpopulation of individuals with target value 0, so the analysis of q_{PS} is closely related to the analysis of $D_{\mathcal{S},n}$. More concretely, if one observes a quality value q_{PS} in a data set and if one wants to know if this value is statistically significantly larger than zero, one can use the Vapnik-Chervonenkis inequalities[1]. Of course, these inequalities are distribution-free and thus very loose, so, if computationally possible, a permutation test would give more information about statistical significance. However, we are interested here in regularization and for this, a permutation test seemingly does not help, because it gives no constructive insight. Therefore we rely here on the Vapnik-Chervonenkis inequalities to firstly analyze the first method of regularization and to secondly guide the second method of regularization.

Remark The Vapnik-Chervonenkis inequalities are obtained by using the union bound and by incorporating the size of the projected family $\mathcal{S}_A$. The growth function and the VC dimension are only instrumental to bounding this size. In our situation, we can always directly bound the size of the projected family by the size of the concept lattice, which is in our case finite and gives the sharpest possible bound, given we assume that the formal context includes only objects which are observed at least once in the sample. But it still seems to be the case that the growth function or the VC dimension are somehow more interesting characteristics in our situation. In Section 3 we will see two families of the same size, but differently behaving growth function, which lead to a very different distribution of the statistic. One reason may be that the size of the concept lattice is only one global characteristic, whereas the growth function and also the VC dimension, which bounds the growth function, contains more 'local information' about 'domains' of V, i.e. $m^{\mathcal{S}}(r)$ does not only tell the size of the projected family for r equals the actually observed sample size, but also for smaller sample sizes/subdomains of V.

[1] Actually, one would have to use inequalities for a two sample situation and one would have to specify the null hypothesis as 'The distribution of the covariates given the target variable are identical and the whole sample is exchangeable with fixed sample sizes for both subpopulations.', otherwise, a conditional analysis, given randomly observed sizes of the subpopulations can be done.

2. Two methods of regularization

We now present two methods of regularization and analyze them w.r.t. their behavior concerning the VC dimension of the regularized subgroup spaces. While the first method is in the spirit of subgroup discovery, the second method is based on ideas of formal concept analysis.

2.1. Method 1: Constraining the description length

One very simple way to reduce $\mathcal{S}$ is to constrain the description length of the considered formal concepts to a maximal length $K \in \mathbb{N}$. We say that a formal concept (A, B) is K-attribute-generated if there exists an attribute set $\tilde{B} \subseteq M$ with $|\tilde{B}| \leq K$ that generates the formal concept (A, B) as $(A, B) = (\tilde{B}', \tilde{B}'')$. In a dual manner, we say that a formal concept is K-object-generated if there exists an object set $\tilde{A} \subseteq G$ with $|\tilde{A}| \leq K$ and $(A, B) = (\tilde{A}'', \tilde{A}')$. Define $\mathcal{S}^K$ as the set of all concept extents of formal concepts that are K-attribute-generated. We now analyze, how the constraining of the description length exactly reduces $\mathcal{S}$ in terms of the VC dimension and in terms of the cardinality $|\mathcal{S}^K|$. For this we introduce a local notion of the VC dimension as follows: For an arbitrary set $A \subseteq G$ define the local VC dimension of $\mathcal{S}$ w.r.t. the location A as the VC dimension of the family $\mathcal{S}_A = \{B \cap A \mid B \in \mathcal{S}\}$. Then we can state the following theorem (cf., also [2, corollary 5.4.5, p.53] for a similar, but still different statement):

Theorem 1.

i) $\mathcal{S}^K \subseteq \mathcal{S}$.
ii) If a domain $A \subseteq G$ does not suffer from a high VC dimension in the weak sense that the local VC dimension of $\mathcal{S}$ w.r.t. A is smaller than or equal to K, then we have $\mathcal{S}_A = \left(\mathcal{S}^K\right)_A$.
iii) The reverse statement of ii) is generally not true.

Proof.
i). This is obvious, because $\mathcal{S}^K$ is a special subset of extents of $\mathbb{K}$, namely that extents, which correspond to K-intent-generated concepts of $\mathbb{K}$.
ii). For arbitrary $A \subseteq G$, because of $i)$ its is clear that $(\mathcal{S}^K)_A \subseteq \mathcal{S}_A$. Consider now the subcontext $\mathbb{K}_A := (A, M, I \cap A \times M)$. Then because of $\mathcal{S}_A = \mathfrak{E}(\mathbb{K}_A)$ the local VC dimension of $\mathcal{S}$ w.r.t. A is exactly the VC dimension of $\mathbb{K}_A$. Let now the VC dimension of $\mathbb{K}_A$ be $D \leq K$. We show that every formal concept of $\mathbb{K}_A$ is K-attribute-generated and thus we also have $\mathcal{S}_A = \mathfrak{E}(\mathbb{K}_A) \subseteq (\mathcal{S}^K)_A$: Let (C, B) be a formal concept of $\mathbb{K}_A$. Look now at the intent B. If $|B| \leq K$ then (C, B) is K-attribute-generated and we are done. If $|B| > K$, then B cannot be shatterable and is therefore not implication-free (w.r.t. $\mathbb{K}_A$), because the VC dimension of $\mathbb{K}_A$ was $\leq K$. We thus have an implication $Y \longrightarrow Z$ with Y and Z nonempty and disjoint and $Y, Z \subseteq B$, which is valid in $\mathbb{K}_A$. This means that with $B_1 := B \backslash Z \subsetneq B$ we have a strict subset of B that generates the same concept $(A, B) = (B_1', B_1'')$. As long as $|B_i| > K$, one can repeat the above argument and obtains further sets $B_1 \supsetneq B_2 \ldots \supsetneq B_j$ until one reaches a subset $\tilde{B} \subsetneq B$

with $|\tilde{B}| \leq K$ that also generates the concept $(A, B) = (\tilde{B}', \tilde{B}'')$. This shows that every concept of $\mathbb{K}_A$ is K-attribute-generated and thus we all-together have $\mathcal{S}_A = \left(\mathcal{S}^K\right)_A$.
iii). We simply give an example of a context with VC dimension 3 where all formal concepts are obviously 1-attribute-generated:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
g_1		x			x	x		x
g_2			x		x		x	x
g_3				x		x	x	x

Statement iii) of Theorem 1 means that not necessarily for every domain A with high local VC dimension the local VC dimension is reduced. Of course, globally the size of $\mathcal{S}$ is reduced, but not directly in terms of the VC dimension. Concretely, one can still obtain an inequality for the size of the regularized family of sets and a generalization inequality, which is very similar to the VC-inequality. If one does look at K-attribute-generated concepts but instead at K-object-generated concepts, then one obtains dual inequalities, which are still more similar to the VC-inequalities:

Theorem 2.

i) The size of the regularized family $\mathcal{S}^K$ is bounded as $|\mathcal{S}^K| \leq \sum_{i=0}^{K} \binom{|M|}{i}$, which gives the generalization inequality $P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot \Gamma(|M|, K) e^{-\frac{n\varepsilon^2}{4}}$.
ii) If we do not constrain the generating attribute sets, but the generating object sets, then for the set $\tilde{\mathcal{S}}^K$ of all K-object-generated concepts we have the bound $|\tilde{\mathcal{S}}^K| \leq \sum_{i=0}^{K} \binom{|G|}{i}$, which gives the generalization inequality $P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot \Gamma(|G|, K) e^{-\frac{n\varepsilon^2}{4}}$, which is identical to the bound one would obtain for a VC dimension K and a sample size $|G|/2$.

Proof. The proof follows directly from that fact that there are not more K-attribute-generated (K-object-generated) formal concepts than attribute descriptions (object descriptions) of size K.

2.2. Method 2a: Identifying large shatterable sets and reducing the set of intents

Another way of regularization would be to firstly directly identify all shatterable object sets of a size larger than K. Let sh_K be the set of all shatterable object sets of size larger than K and let $N := \bigcup sh_K$ be the set of all objects that belong to some shatterable set of size larger than K. Define $\mathbb{K}_{\setminus N} := (G \setminus N, M, I \cap G \setminus N \times M)$ as the context restricted to the objects which do not belong to shatterable sets of size larger than K. Define further $\mathcal{T}^K := \{\{g \in G \mid \forall m \in B : gIm\} \mid B \text{ intent of } \mathbb{K}_{\setminus N}\}$ as the set of all concept extents of $\mathbb{K}$ which are generated by intents of the subcontext $\mathbb{K}_{\setminus N}$. Then we have the following

Theorem 3.

i) If a domain $A \subseteq G$ does not suffer from a high VC dimension in the strong sense that $A \subseteq G\backslash N$ (which implis that $\mathcal{S}_A$ has a VC dimension of at most K, but not vice versa), then $\mathcal{S}_{\mathcal{A}} = \left(\mathcal{T}^K\right)_A$.
ii) $\mathcal{T}^K \subseteq \mathcal{S}$.
ii) $(\mathcal{T}^K, \subseteq) \cong (\mathfrak{E}(K_{\backslash N}), \subseteq)$, i.e. the family $(\mathcal{T}^K)$, equipped with the subset-relation, is order-theoretically isomorphic to the family of the extents of the reduced context $\mathbb{K}_{\backslash N}$, equipped with the subset-relation.
iv) Because of iii) we have $|\mathcal{T}^K| = |\mathfrak{E}(\mathbb{K}_{\backslash N})| \leq \Gamma(min(|G\backslash N|, |M|), K)$. (Note that the VC dimension of $\mathfrak{E}(\mathbb{K}_{\backslash N})$ is at most K.)
v) The VC dimension of $\mathcal{T}^K$ is at least the VC dimension of $\mathfrak{E}(\mathbb{K}_{\backslash N})$.
vi) Because of iv) we have $P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot \Gamma(min(|G\backslash N|, |M|)|, K) e^{-\frac{n\varepsilon^2}{4}}$.

Remark Of course, due to its construction the family $\mathfrak{E}(\mathbb{K}_{\backslash N})$ has a VC dimension of at most K. But it can still be the case that the order-theoretically isomorphic family $\mathcal{T}^K$ has a VC dimension larger than K. This underlines the fact that the VC dimension is no purely order -theoretic notion, but a combinatorial notion. The formal concepts of $\mathfrak{E}(\mathbb{K})$ and the elements of $\mathcal{T}^K$ are in a one to one relation, thus the families have the same size. But, comparing both families, the temporarily deleted objects, which play also an important role for the VC dimension, do not play any role w.r.t. $\mathfrak{E}(\mathbb{K})$, but in contrast they belong to usually many different sets of $\mathcal{T}^K$, which makes $\mathcal{T}^K$ more complex. The following example shall illustrate this: For the context given in Table 1 and $K = 2$ we have $N = \{g_1, g_2, \ldots, g_6\}$ and the VC dimension of $\mathbb{K}_{\backslash N}$ is 2, which can also be seen in the concept lattice given in the top of Figure 1. The whole context has VC dimension 3, because the objects g_7, g_8 and g_9, together with the attributes m_1, m_2 and m_3 are building a contranominal scale. The temporarily deleted objects g_7, g_8 and g_9 are reincorporated into the concept lattice of $\mathbb{K}_{\backslash N}$ like given at the bottom of Figure 1. Here one can easily see that the object set $\{g_7, g_8, g_9\}$ is shatterable (w.r.t. $\mathcal{T}^K$). The reason is that these objects are associated to more than one concept of the subcontext $\mathbb{K}_{\backslash N}$ in such a way that they become shatterable.

TABLE 1. *Formal context with VC dimension 3 and subcontext with VC dimension 2.*

	m_1	m_2	m_3	m_4	m_5	m_6
g_1						x
g_2	x					
g_3		x				
g_4				x		
g_5					x	
g_6			x			
g_7		x	x			x
g_8	x		x	x		
g_9	x	x			x	

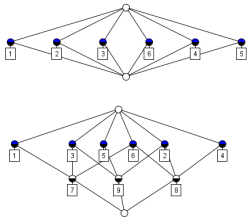

FIG 1. *Formal concept lattice of the subcontext $\mathbb{K}_{\setminus N}$ with VC dimension 2 (top), as well as $\mathcal{T}^K$ (bottom) with reincorporated objects g_7, g_8, g_9 and VC dimension 3.*

2.3. Method 2b: Reducing the context by identifying large shatterable sets and restructuring the original context without increasing the reduced VC dimension

In this section we now want to circumvent the fact that reincorporating the temporarily deleted objects may increase the VC dimension. A useful observation is here that if one associates the temporarily deleted objects not to many concepts, but only to one concept of the subcontext, then the VC dimension will not increase. The reason for this is stated in the following

Lemma 1 Let $\mathbb{K} = (G, M, I)$ be a formal context and let $\mathfrak{I}(\mathbb{K})$ be the family of all intents of this context. Define the associated context $\mathbb{K}^* := (\mathfrak{I}(\mathbb{K}), M, \ni)$. Note that for $g \in G, m \in M$, we have $gIm \iff \{g\}' \ni m$, thus there is an embedding of the objects $g \in G$ into $\mathfrak{I}(\mathbb{K})$ and $\mathbb{K}^*$ is therefore some kind of an extension of $\mathbb{K}$. Note further that the concept lattices of $\mathbb{K}^*$ and $\mathbb{K}$ are order-theoretically isomorphic. Now, beyond this order-theoretical isomorphy, the families $\mathfrak{E}(\mathbb{K}^*)$ and $\mathfrak{E}(\mathbb{K})$ have the same VC dimension.

Proof. The statement directly follows from the fact that the attribute-implications that are valid in $\mathbb{K}^*$ are exactly the attribute-implications that are valid in $\mathbb{K}$, because 'adding' intents of $\mathbb{K}$ to obtain $\mathbb{K}^*$ does not change the validity of implications, because all intents of $\mathbb{K}$ respect all implications that are valid in $\mathbb{K}$.

Lemma 1 allows us to incorporate the temporarily deleted objects by identifying them with special intents of the subcontext without increasing the VC dimension. With which intent one exactly identifies an object does not play a role for the VC dimension, as long as every object is only identified with only one intent. Of course, the objects should be identified in a somehow reasonable manner. One possibility would be to define some sort of a metric between intents and identify an object g with that intent that is closest to $\{g\}'$. Now, we concretize the ideas from above in

Theorem 4. Let $\mathbb{K} = (G, M, I)$ be a formal context and let $\tilde{K} = (\tilde{G}, M, \tilde{I})$ with $\tilde{G} \subseteq G$ and $\tilde{I} = I \cap \tilde{G} \times M$ be a subcontext of $\mathbb{K}$ with VC dimension K. (Think of $\tilde{G} = G \backslash N$.) Let furthermore be $R : G \longrightarrow \mathfrak{I}(\tilde{\mathbb{K}})$ an arbitrary mapping. (Think of a mapping R which satisfies $R(g) = \{g\}'$ for $g \in \tilde{G}$ and that $R(g)$ is in some certain sense close to $\{g\}'$ for $g \in G \backslash \tilde{G}$.) Define the restructured family $\mathcal{U}^R$ of sets as $\mathcal{U}^R := \{\Phi^*(B) \mid B \in \mathfrak{I}(\tilde{\mathbb{K}})\} = \{\Phi^*(B) \mid B \subseteq M\}$ with $\Phi^* : 2^M \longrightarrow 2^G : B \mapsto \{g \in G \mid \forall m \in B : R(g) \ni m\} = \{g \in G \mid \forall m \in B : gI^*m\}$ and $gI^*m :\Longleftrightarrow\ m \in R(g)$. Then we have

i) If a domain $A \subseteq G$ does not suffer from a high VC dimension in the strong sense that $A \subseteq \tilde{G}$, then $\mathcal{S}_{\mathcal{A}} = \left(\mathcal{U}^R\right)_A$.
ii) In general, it does not hold that $\mathcal{U}^R \subseteq \mathcal{S}$.
iii) But of course it holds that $|\mathcal{U}^R| = |\mathfrak{E}(\tilde{K})| \leq |\mathcal{S}|$, and thus we have the generalization inequality

$$P^{\otimes n}\left(D_{\mathcal{S},n} \geq \varepsilon\right) \leq 6 \cdot \Gamma(min(|G\backslash N|, |M|)|, K) e^{-\frac{n\varepsilon^2}{4}}.$$

1. The VC dimension of $\mathcal{U}^R$ is at most K.

3. Analyzing the methods of regularization by means of a dataset from the German General Social Survey (GGSS)

In this section we briefly analyze a data set from the German General Social Survey 2018 ([1]). We use the ISSP 2017 module on social networks and social resources. As covariates/attributes for the analyzed context we use the answers to 10 questions. The questions were about if the respondents did know several people who have a job from a list of 10 occupational groups (e.g., bus/lorry driver). The possible answers were: 'Family or relative', 'Close friend', 'Someone else I know' or 'No one'. (A fifth answer category 'Can't choose' was disregarded for the analysis.) These 10 nominal variables with four possible outcomes were nominally conceptually scaled, which leads to $|M| = 40$ attributes. All-together 1354 respondents answered the questions, thus, we have $|G| = 1354$. As a target variable we use the answer to the question 'At these occasions [going out to eat or drink with three or more friends or acquaintances who are not family members], how often do you make new friends or acquaintances?'. We dichotomized the target variable as being 1 if the answer was 'often' or 'very often' and 0 otherwise. The subgroup with the highest value of the Piatetsky-Shapiro quality function was the subgroup of respondents who know a hairdresser/barber as a close friend ($q_{PS} \approx 10.19$ and $D := \sup_{A \in \mathcal{S}}(\nu^1(A) - \nu^2(A)) \approx 0.154$). Based on a permutation test, the statistic was not significantly positive ($p \approx 0.15$). The original, non-regularized family was explicitly computable and consisted of 81622 formal concepts and had a VC dimension of 10. The set N for Method 2 was computed by identifying for each of the 1354 objects the largest contranominal scale that contains the corresponding object. This was done using a mixed integer linear programming formulation of the problem. For the mapping R of method 2b we used a metric approach with $R(g) \in \text{argmin}\{d(\{g\}', I) \mid I \in \mathfrak{I}, I \subseteq \{g\}'\}$ and

d the Manhattan metric, weighted by the column sums of the subcontext. To analyze the actual behavior of the different regularization methods, for each family $\mathcal{S}$, we randomly draw subsets $A \subseteq G$ and plotted the size $|\mathcal{S}_A|$ of the projected family against the size of A. This gives a very rough insight into the progression of the growth function. For the family $\mathcal{T}^K$ we considered the set of all 3-object-generated concepts, i.e., $K = 3$. For the second method of regularization, we choosed $K = 6$. The reason for this difference in K was to make the corresponding families of sets more comparable. The exact choice was based on the (of course somehow arbitrary) requirement that the families of sets are comparable in its size. The families were not too large, so we could compute them explicitly, concretely we had $|\mathcal{S}| = 8261$; $|\mathcal{S}^2| = 762$; $|\mathcal{S}^4| = 37937$ and $\mathcal{T}^7 = 42978$; $|\mathcal{T}| = 3299$. For the set N of method 2 we decided to take the set of all objects belonging to a contranominal scale of maximal size 6 plus some of the objects that belong to contranominal scales of maximal size 7, based on the number of contranominal scales of size 7 to which they belong to. With this method of breaking ties, we got a family of size 8248, which is approximately comparable to the size 8261 for the first method. From Figure 2 (left) it appears that the growth function for method 1 is the most concave one and method 2b has the least concave growth function. The growth function of method 2b appears to be approximately linear in the double-logarithmic plot, which corresponds to a polynomial growth. For domains of medium size, method 2b seems to regularize the subgroup space most strongly compared to the other methods and relative to the regularization of large domains. Note that we somehow arbitrarily matched the strength of the regularization of the methods by matching the size of the regularized families. Whatever that means and which behavior is most desirable, the methods have a clear distinguished behavior concerning the growth function. The right side of Figure 2 shows the behavior of the statistic D under a simulation by permuting the target variable, which corresponds to a null hypothesis that there is no difference between the subpopulations with different target values concerning the distribution of the covariates. Under the null hypothesis D should be stochastically small to make a test against the null very sensitive. One can see that the distribution for method 1 and for method 2a are very similar, but for method 2a the distribution is slightly stochastically smaller. (The distribution of the non-regularized statistic is more or less the same as for method 1, in this sense, method 1 does not really regularize the statistic at all, so a smaller K would be needed for method 1.) In comparison, for method $2a$ the distribution of D is clearly stochastically smaller than for the other methods, actually with this regularization, the result becomes statistically significant in the sense of $p \approx 0.02$ (for method 2a we have $p \approx 0.14$). One important thing to note is that all methods have nearly the same size, which leads to nearly the same generalization bound, but w.r.t. the distribution of the test statistic they behave differently, which is maybe due to the differently behaving growth function. Of course, this is somehow speculative. Additionally, for method 2b the value of the statistic for the actually observed data is here slightly larger than the value for the non-regularized statistic ($D = 0.160$, $q_{PS} = 10.60$, for the other methods, the value did not change). Note that this

cannot happen for method 1 and method 2a since the families are subfamilies of the non-regularized families. So, the restructuring happened here 'in the right direction' in the sense of making D smaller under the null and larger for the actually observed data, but this could also be somehow accidental.

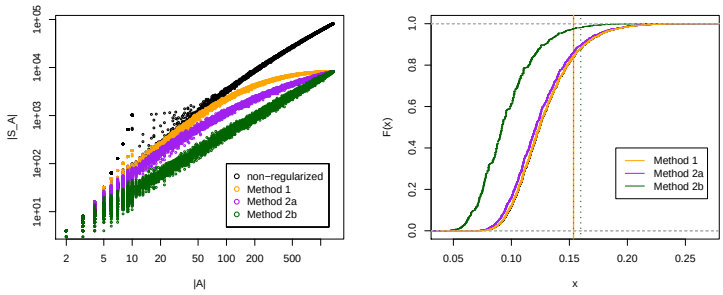

FIG 2. *simulation-based estimate of the growth function (left) and the distribution of the statistic D under H_0 based on a permutation scheme (right).*

4. Conclusion

In this paper we analyzed two different methods of regularization for subgroup discovery. We worked out some analytic results on the size and the VC dimension of the regularized subgroup space and presented some contingent differences concerning the growth function that appeared in a data example. Questions of further research are firstly: Can one establish criteria which could shed more light on the interplay between the size and the growth function/VC dimension beyond the VC inequalities in the sense of establishing criteria which could somehow guide, which characteristic to control for regularization in which cases? Secondly, our method 2b used a restructuring of the subgroup space. This is somehow non-typical compared to for example the methodology of structural risk minimization where one always has a nested structure of families. This imposes the question, if one can establish criteria, which could shed light on the relationship between the method of reducing a family in size versus restructuring a family? Thirdly, one can ask, which mappings for restructuring the family could be useful?

References

[1] Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2018. GESIS Datenarchiv, Köln. *GESIS - Leibniz-Institut für Sozialwissenschaften (2019).*

[2] A. L. J. H. Albano. *Polynomial Growth of Concept Lattices, Canonical Bases and Generators: Extremal Set Theory in Formal Concept Analysis.* PhD thesis, Saechsische Landesbibliothek-Staats-und Universitaetsbibliothek Dresden, 2017.

[3] W. Duivesteijn and A. Knobbe. Exploiting false discoveries – statistical validation of patterns and quality measures in subgroup discovery. In *2011 IEEE 11th International Conference on Data Mining*, pages 151–160, 2011.

[4] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–238, 1991.

[5] V. N. Vapnik and A. Y. Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

[6] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[7] G. I. Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.

[8] S. Wrobel. Inductive logic programming for knowledge discovery in databases. In *Relational data mining*, pages 74–101. Springer, 2001.