Novel approach for phenotyping based on diverse top-k subgroup lists

Antonio Lopez-Martinez-Carrasco[1(✉)][0000−0002−2990−886X], Hugo M. Proença[2][0000−0001−7315−5925], Jose M. Juarez[1][0000−0003−1776−1992], Matthijs van Leeuwen[2][0000−0002−0510−3549], and Manuel Campos[1,3][0000−0002−5233−3769]

[1] MedAI-Lab, University of Murcia, Spain
[2] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
[3] Murcian Bio-Health Institute (IMIB-Arrixaca), Spain

Abstract. The discovery of phenotypes is useful to describe a population. Providing a set of diverse patient phenotypes with the same medical condition may help clinicians to understand it. In this paper, we approach this problem by defining the technical task of mining diverse top-k phenotypes and proposing an algorithm called DSLM to solve it. The phenotypes obtained are evaluated according to their quality and predictive capacity in a bacterial infection problem.

Keywords: Patient phenotyping · Subgroup Discovery · Subgroup List · the Minimum Description Length principle · Algorithm.

1 Introduction

Phenotyping consists of finding a set of observable characteristics of patients with a medical condition [5]. Unfortunately, getting a single explanation of a medical phenomenon might be poor under the eyes of a clinical researcher, and observable descriptions could be ruled out. For this reason, multiple phenotypes might be provided, which implies that both the characteristics used and the patients represented must be diverse, thus providing the clinician with different explanations for the same medical phenomenon. From a machine learning perspective, phenotyping could be interpreted as the automatic generation of models describing a subset of the population with a specific feature. Therefore, we propose to use the Subgroup Discovery (SD) paradigm [1] as a building block of phenotypes.Moreover, the generation of multiple subgroup list models [3, 4] is a suitable approach to generate diverse phenotypes. From a clinical perspective, a critical problem when creating subgroup lists is the large number of candidate subgroups that are mined by the SD algorithm. A feasible solution is to adopt the Minimum Description Length (MDL) principle, which is a method of inductive inference based on the idea that the model that best explains the data is the one that best compresses the data.

The contributions of this work are: (1) the definition of the new problem of mining diverse top-k phenotypes, and (2) a new algorithm called DSLM that uses

SD and the MDL principle to solve this problem. These contributions provide clinicians with a method to obtain multiple and diverse explanations of a set of patients with statistically robust subgroups.

2 Problem definition and background

This section formally defines the novel problem of mining diverse top-k phenotypes and provides a short background of the methods used.

The following basic concepts are first introduced: (1) an attribute a is a relation between an object property and its value; (2) the domain of a ($dom(a)$) is the set of all unique values that a can take; (3) an instance i is a tuple $i = (a_1, \ldots, a_m)$ of attributes; and (4) a dataset d is a tuple $d = (i_1, \ldots, i_n)$ of instances. Note that an attribute can be nominal or numeric depending on its domain and that the notation $v_{x,y}$ is used to indicate the value of the x-th instance i_x and its y-th attribute a_y from a dataset d. Subsequently, the following definitions can be given. Given an attribute a_y from a dataset d, a binary $operator \in \{=, \neq, <, >, \leq, \geq\}$ and a value $w \in dom(a_y)$, then a selector e is defined as a 3-tuple of the form $(a_y, operator, w)$. In our problem, descriptions contained in a phenotype are related to a specific target value represented by a selector e. Given an instance i_x and a selector $e = (a_y, operator, w \in dom(a_y))$, then $selector\_coverage(i, e)$ function returns 1 if the binary expression "$v_{x,y}$ $operator$ w" holds $true$, and returns 0 otherwise. A pattern p is a set of selectors of the form $\{e_1, \ldots, e_j\}$ that represents a conjunction of conditions describing a subset of the dataset, and in which all attributes of the selectors are different. Given an instance i and a pattern p, then $pattern\_coverage(i, p)$ function returns 1 if $\forall e_j \in p, selector\_coverage(i, e_j)$, and returns 0 otherwise. Additionally, we say that a pair (p, e) is positive if it exists an instance i covered by both p and e.

Given a dataset d and a target e, then a phenotype l is a list of patterns $< p_1, \ldots, p_z >$ such that $\forall p_z \in l, (p_z, e)$ is positive, and that cover dataset instances that are statistically different and interesting, compared to the dataset distribution. In the problem defined, we consider that multiple phenotypes must contain positive pairs pattern-target since the objective in this domain is to explain a specific target value from a dataset (e.g., $exitus = yes$).Moreover, when ensuring diversity, the outcome must provide clinicians with multiple phenotypes that are different and non-redundant. Finally, the problem of mining diverse top-k phenotypes is defined as follows: given a dataset d, a target e and the k maximum number of phenotypes to discover, the problem of mining diverse top-k phenotypes consists of generating a set of phenotypes $\{l_1, \ldots, l_k\}$ such that they only contain positive pairs pattern-target and are diverse.

Given a pattern p and a selector e, a subgroup s is a pair (p, e) in which the pattern is denominated as 'description' and the selector is denominated as 'target'. Additionally, given a subgroup s and a dataset d, a quality measure q is a function that computes one numeric value according to s and certain metrics from d (e.g., WRAcc). Finally, the SD problem consists of exploring the search

space of a dataset d to mine subgroups whose quality value q is greater or equal to a given *threshold*.

A subgroup list [4] is a collection of ordered subgroups followed by a default subgroup, whose objective is to iteratively divide the input data into different subsets and to provide a description for each of them, except the last one, which corresponds to the default subgroup. While subgroups contained in a subgroup list cover the instances that are statistically different in comparison with the dataset distribution, the default subgroup represents the dataset average, covering the instances that are well described by the dataset distribution.

The authors of [4] used the MDL principle to build a single subgroup list. They defined the MDL encoding of the optimal subgroup list for a certain dataset and proposed a greedy algorithm called SDD++ that iteratively added one subgroup at a time to the subgroup list.

3 DSLM algorithm

Diverse Subgroup Lists Miner[4] (DSLM) aims to generate diverse top-k subgroup lists by using SD and the MDL principle. This algorithm requires the following inputs: a dataset d, a collection $\mathcal{C}$ of subgroups, the number k of subgroup lists to generate, the maximum number of subgroups for each subgroup list (sl_max_size), and the maximum overlap permitted ($max\_overlap$). Note that subgroups contained in $\mathcal{C}$ can be generated by any SD algorithm and later filtered before executing this algorithm.

DSLM algorithm initially creates an empty collection $\mathcal{L}$. Next, it iterates k times and, in each loop, initializes an empty subgroup list, adds it to $\mathcal{L}$ and iterates over $\mathcal{C}$ to fill it. For each candidate from $\mathcal{C}$, it is selected only if: (1) its score according to the MDL principle and the overlap factor is higher than the best so far, (2) it is positive (i.e., the pair pattern-target by which it is formed is positive), and (3) its overlap with the subgroups already contained in the subgroup list is less or equal to the maximum permitted. The overlap is controlled by the counter list that is obtained by $compute\_overlap\_factor$ function. This function computes the proportion between the number of instances covered by the candidate c and the number of instances covered by the subgroup list sl (stored in a list called $overlap\_counter$). Note that, for computing the overlap, subgroups are considered individually (i.e., not depending on their position in the subgroup list). When the best candidate is selected: (1) overlap counter is updated, (2) it is added to the subgroup list, (3) its refinements are deleted of $\mathcal{C}$, and (4) all subgroups from $\mathcal{C}$ of which it is a refinement are deleted. Note that both the overlap computation and these last operations contribute to diversity. What is more, they also reduce the number of candidates to explore, which would not be selected anyway due to their overlap. Finally, the collection $\mathcal{L}$ with top-k subgroup lists is returned.

[4] Available at `subgroups` python library or https://github.com/antoniolopezmc/subgroups

4 Experiments and Discussion

The experiments aim to validate our proposal in the context of phenotyping antimicrobial resistances. We used real clinical data extracted from MIMIC-III public database using as target the patients infected by an Enteroccous Sp. bacterium resistant to Vancomycin. The final dataset had 9,240 instances and 12 attributes. Then, VLSD [2] exhaustive SD algorithm was applied (using WRAcc quality measure with a threshold of 0 and a maximum depth of 3), mining 473 subgroups. After that, the DSLM algorithm was executed using the previous subgroups and with $k = 2$, $sl\_max\_size = 6$ and $max\_overlap = 0.06$, obtaining the subgroup lists depicted in Table 1.

	Subgroup description (Pattern)	Subgroup coverage Pos-Neg	Contribution Pos-Neg	Phenotype coverage Pos-Neg
s1	culture_type = 'SWAB', icu_when_culture = 'SICU'	466-168	466-168	466-168
s2	culture_type = 'URINE', previous_vancomycin = 'yes'	145-85	145-85	611-253
s3	days_admitted_before_ICU = 'OneDayOrMore', discharge_location = 'DEAD/EXPIRED', patient_age = 'ADULT'	167-100	147-96	758-349
s4	culture_type = 'BLOOD_CULTURE', previous_vancomycin = 'yes'	112-111	98-110	856-459
s5	admission_location = 'PHYS_REFERRAL/NORMAL_DELI', readmission = 'yes', service_when_culture = 'SURG'	124-86	81-80	937-539
s6	culture_type = 'SWAB', previous_vancomycin = 'yes'	114-97	76-89	1013-628
	Default subgroup	-	1113-6486	2126-7114

	Subgroup description (Pattern)	Subgroup coverage Pos-Neg	Contribution Pos-Neg	Phenotype coverage Pos-Neg
s1	culture_type = 'SWAB', service_when_culture = 'SURG'	380-241	380-241	380-241
s2	days_admitted_before_ICU = 'OneDayOrMore', previous_vancomycin = 'yes'	166-136	150-126	530-367
s3	service_when_culture = 'OMED'	115-95	85-92	615-459
s4	days_admitted_before_ICU = 'ZeroDays', previous_vancomycin = 'yes', service_when_culture = 'SURG'	110-77	76-64	691-523
s5	culture_type = 'SWAB', service_when_culture = 'MED'	205-321	193-320	884-843
s6	days_admitted_before_ICU = 'OneDayOrMore', discharge_location = 'DISTINCT_PART_HOSP', service_when_culture = 'SURG'	125-121	76-81	960-924
	Default subgroup	-	1166-6190	2126-7114

Table 1. Top-2 phenotypes from our dataset.

The first phenotype describes patients admitted to an ICU ward, adults, and different types of culture (swab, urine, or blood), with treatment by vancomycin in previous admissions, and that could die. The second phenotype represents patients from emergencies or medical wards (SURG, OMED or MED), that could or not be complicated and moved to ICU, and the cultures were swab mainly. Moreover, it also had vancomycin in previous admissions. Finally, to evaluate the phenotypes from an objective point of view, we show their predictive capacity by using them as dummy variables in a classification algorithm that predicts the target variable of the phenotype. Two classification models were fitted: (1) with the original dataset, and (2) adding two attributes to the original dataset, indicating whether an instance is covered or not by each phenotype. Random

Forest, Gradient Boosting Classifier and Logistic Regression were used and the second classification model always obtained a statistically higher accuracy than the first one, demonstrating that both phenotypes have predictive capacity.

5 Conclusions

In this research, we proposed a novel approach for phenotyping based on the task of mining diverse top-k subgroup lists. The aim was to provide clinicians with few descriptions about a specific target value of interest that are diverse both in coverage and in descriptions. Moreover, we also proposed the Diverse Subgroup Lists Miner algorithm (DSLM) algorithm, which generates subgroup lists based on the subgroup discovery paradigm and the minimum description length principle. We carried out the experiments about phenotyping antimicrobial resistances in the MIMIC-III database. The results showed that the top-2 phenotypes represented by the subgroup list model have valuable properties: they are statistically robust, they are legible by the clinical experts, they are diverse, and they have few selectors. Finally, the predictive capacity of the phenotypes obtained has been probed by significantly increasing the accuracy of all the classification algorithms used to predict the same outcome of the phenotype after including the phenotypes as new independent variables in the dataset.

Acknowledgements: this work was partially funded by the CONFAINCE project (Ref: PID2021-122194OB-I00) by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union", and by the GRALENIA project (Ref: 2021/C005/00150055) supported by the Spanish Ministry of Economic Affairs and Digital Transformation, the Spanish Secretariat of State for Digitization and Artificial Intelligence, Red.es and by the NextGenerationEU funding. This research was also partially funded by a national grant (Ref: FPU18/02220), of the Spanish Ministry of Science, Innovation and Universities (MCIU) and by a mobility grant (Ref: R-933/2021), of the University of Murcia.

References

1. Lavrac, N., Kavsek, B., Flach, P.A., Todorovski, L.: Subgroup Discovery with CN2-SD. J. Mach. Learn. Res. **5**, 153–188 (2004)
2. Lopez-Martinez-Carrasco, A., Juarez, J.M., Campos, M., Canovas-Segura, B.: VLSD - An efficient Subgroup Discovery algorithm based on Equivalence Classes and Optimistic Estimate. Knowledge and Information Systems. Status: submitted.
3. Lopez-Martinez-Carrasco, A., Proença, H.M., Juarez, J.M., van Leeuwen, M., Campos, M.: Discovering diverse top-k characteristic lists. In: 21th Symposium on Intelligent Data Analysis (IDA 2023) (2023)
4. Proença, H.M., Grünwald, P., Bäck, T., van Leeuwen, M.: Robust subgroup discovery. Data Mining and Knowledge Discovery (2022)
5. Wojczynski, M.K., Tiwari, H.K.: Definition of phenotype. In: Genetic Dissection of Complex Traits, Adv. in Genetics, vol. 60, pp. 75–105 (2008)