Subjectively Interesting Subgroup Discovery on Real-valued Targets

Jefrey Lijffijt*, Bo Kang*, Wouter Duivesteijn+, Kai Puolamäki=#, Emilia Oikarinen=#, Tijl De Bie*

* *Department of Electronics and Information Systems, IDLab, Ghent University*
+ *Department of Computer Science and Mathematics, Eindhoven University of Technology*
= *Department of Computer Science, Aalto University*
Finnish Institute of Occupational Health

***Abstract*—Deriving insights from high-dimensional data is one of the core problems in data mining. The difficulty mainly stems from the large number of variable combinations to potentially consider. Hence, an obvious question is whether we can automate the search for interesting patterns. Here, we consider the setting where a user wants to *learn* as efficiently as possible about real-valued attributes. We introduce a method to find subgroups in the data that are maximally informative (in the Information Theoretic sense) with respect to one or more real-valued target attributes. The succinct subgroup descriptions are in terms of arbitrarily-typed description attributes. The approach is based on the Subjective Interestingness framework FORSIED to use prior knowledge when mining most informative patterns.**

I. Introduction

We introduce the central ideas of this paper by means of an example. Consider a user who wants to learn about crime demographics from the UCI Communities and Crime data. This data contains violent crime rates for all ($n = 1994$) districts in the USA, and over 120 other attributes describing demographic statistics of those districts. One method to learn about relations between the 'number of violent crimes' attribute and the demographic attributes is to extract *subgroup patterns*. These are sets of data points where violent crime is surprisingly high (or low), and that can be succinctly described in terms of intervals over one or several demographic attributes. A subgroup pattern should be interpreted as 'for data points that fall within the specified statistics that describe the subgroup, violent crime is surprisingly low/high'.

For example, the top subgroup pattern—identified through the method introduced in this paper—states that there are high violent crime rates in districts where many mothers are unmarried at the moment they give birth to their child (condition $PctIlleg >= 0.39$; mean violent crime rate 0.53 in subgroup vs. 0.24 overall). An illustration of the data coverage for this pattern is given in Fig. 1. The subgroup covers 20.5% of the data and may be interesting because the distribution of crime rates within this subgroup deviates substantially from the full data. If a user would have no prior expectations about the data, this pattern is highly informative.

Indeed, we may quantify how informative/interesting the pattern is, in the Information Theoretic sense: the number of bits of information we gain about the data by learning about this pattern, which depends on the amount of data covered (more is better) and how much the distribution in the subgroup differs from our expectation (also larger is better). In this paper, we consider mean and variance statistics. Typically, we would like to weight this information gain against how complex the description of the pattern is (a function of number of attributes used to describe the subgroup plus the number of statistics presented to the user; smaller is better), such that our aim is to provide a maximal *information rate*.

This is precisely the contribution of this paper. We quantify the Information Content (IC; the amount of information gained) and Description Length (DL; the complexity of the description) for *subgroup patterns* for the case of first and second order statistics. While the example above has one *target attribute* (the violent crime rate), we also do this for target sets, to enable users to learn about mean and (co-)variance statistics of multivariate distributions. This also allows discovery of subgroups with surprising *interactions* between targets, a concept known as Exceptional Model Mining.

As hinted at in the example, the IC of a pattern is inherently *subjective*. That is, it is particular to a user, because *how much you learn depends on your prior knowledge*. We implement this subjectivity by modeling a background distribution over the data space; a Maximum Entropy distribution subject to constraints corresponding to the current knowledge of a user. This approach is known as FORSIED [1]. It also provides principles to enable iterative mining of non-redundant patterns without much additional effort.

We have implemented an algorithm to iteratively mine interesting patterns which is freely available as open source

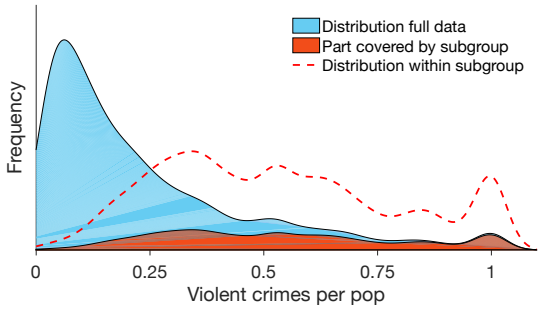

Fig. 1. Distribution of violent crime over the full data (light blue area), part covered by the subgroup 'high rate of unmarried mothers' (red area), and distribution within the subgroup (red dotted line). Height of colored areas given by Gaussian-kernel smoothed estimates. The subgroup clearly covers a substantial amount of the data where the violent crime rate is relatively high.

code. We have not studied the algorithmic problem in detail, but the implementation is based on beam search, a frequently employed approach in subgroup discovery. That is, it maintains a list of most interesting patterns of arity k, expands these to arity $k+1$ and selects the most interesting patterns again. Ultimately, it outputs the most interesting pattern found. It handles categorical, ordinal, and numerical *description attributes* (the demographic attributes in the example) and supports time constraints (e.g., stop after 1 minute of mining). The implementation is based on Cortana [2].

In summary, this paper contributes the following:

– We present a new pattern syntax for subgroups in multivariate real-valued data, called *location* and *spread* patterns.
– We summarize how to quantify their interestingness in a subjective manner, and how to incorporate prior knowledge and previously mined patterns into a background model used in the interestingness score, to mine non-redundant patterns.
– We discuss how to algorithmically find high-quality patterns.
– We show experimental results on several datasets.

This paper is a summary of the technical report [3]. Discussion of related work can be found in the extended report. All code is available at: https://bitbucket.org/ghentdatascience/sisd-public/.

II. Methods

Overview. The high-level problem addressed in this paper is: *Iteratively inform the user about subsets of data points that can be described concisely and that have surprising mean or variance statistics, such that the rate of information gain of the user about the target attributes is maximized at each iteration.*

Our formalization of the problem follows the FORSIED approach: we model a background distribution over the data space that represents the user's (evolving) belief state and we quantify the IC of a pattern as the information the user gains about the target attributes by seeing the pattern.

Notation. Let the data be a set of n pairs $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)$, $i \in \{1, \ldots, n\}$, where the *description attributes* of the ith data point $\hat{\mathbf{x}}_i \in \prod_{j=1:d_\mathbf{x}} \mathcal{X}_j$ are a tuple of $d_\mathbf{x}$ attributes with domains $\mathcal{X}_j$, and $\hat{\mathbf{y}}_i \in \mathbb{R}^{d_\mathbf{y}}$ is a vector containing the values for $d_\mathbf{y}$ real-valued *target attributes*. We denote $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}'_1, \hat{\mathbf{y}}'_2, \cdots, \hat{\mathbf{y}}'_n)'$. Hatted symbols indicate empirical values and non-hatted the respective random variables.

Subgroups, intentions, and extensions. A subgroup is defined by a set of *conditions* on the description attributes (the value combination is the *intention*), together with the set of data points whose description attributes satisfy the conditions (the index set $\mathcal{I} \subseteq \{1, \ldots, n\}$ is the subgroup *extension*).

Location and spread patterns. Subgroups tend to be informative if the target attribute values of data points in the extension $\{\hat{\mathbf{y}}_i | i \in \mathcal{I}\}$ are *unusual*, quantified by means of statistics—functions of this set of data points. We define two statistics $f_\mathcal{I} : \mathbb{R}^{n \times d_\mathbf{y}} \mapsto \mathbb{R}^{d_\mathbf{y}}$ and $g^\mathbf{w}_\mathcal{I} : \mathbb{R}^{n \times d_\mathbf{y}} \mapsto \mathbb{R}$ as follows:

$$f_\mathcal{I}(\mathbf{Y}) = \sum\nolimits_{i \in \mathcal{I}} \mathbf{y}_i / |\mathcal{I}|, \text{ and} \quad (1)$$

$$g^\mathbf{w}_\mathcal{I}(\mathbf{Y}) = \sum\nolimits_{i \in \mathcal{I}} \left((\mathbf{y}_i - \hat{\mathbf{y}}_\mathcal{I})' \mathbf{w} \right)^2 / |\mathcal{I}|, \quad (2)$$

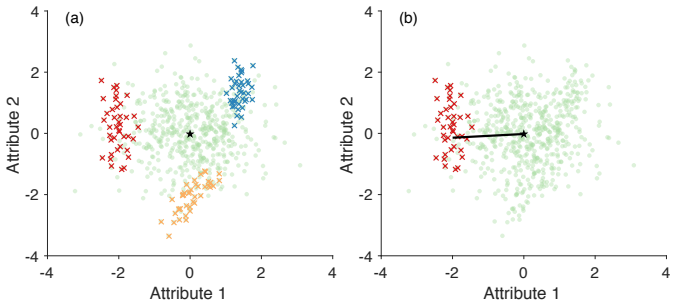

Fig. 2. Patterns found in the synthetic data (§II,§III), (a) Data with the embedded patterns highlighted. (b) Top ranked pattern. Light green circles are random data points, darker colored crosses the three embedded clusters.

where $\hat{\mathbf{y}}_\mathcal{I} = \sum_{i \in \mathcal{I}} \hat{\mathbf{y}}_i / |\mathcal{I}|$ and $\mathbf{w} \in \mathbb{R}^k$ is a unit vector, i.e., $\mathbf{w}'\mathbf{w} = 1$. The first set of $d_\mathbf{y}$ statistics quantifies the average vector of the data points in the extension (i.e., its average *location*), whereas the second quantifies the *spread* around that location, in the direction $\mathbf{w}$. Patterns considered here are specified by an intention, which determines the extension $\mathcal{I}$, a unit vector $\mathbf{w}$, and the specification of the empirical values of one or both $f_\mathcal{I}(\hat{\mathbf{Y}})$ and $g^\mathbf{w}_\mathcal{I}(\hat{\mathbf{Y}})$. These are *location* and *spread* patterns. We specify a direction $\mathbf{w}$ because reading a full co-variance matrix is difficult and very time consuming.

Example. For the synthetic data shown in Fig 2a, a location pattern is an intention, e.g., 'Attribute3 = true', along with the mean of the subgroup, e.g., the dark red set of points. A spread pattern is an intention, a direction (a weight vector, as in Fig. 2b), and the magnitude of the variance in that direction.

Subjective interestingness (SI) quantification. Due to lack of space, we only summarize the approach to quantify the SI of location and spread patterns. For more detail see [3].

We compute a background distribution $p(\mathbf{Y})$ as the Maximum Entropy distribution subject to constraints. These constraints represent the prior knowledge of the user. After showing the user a pattern, we also incorporate that pattern as an additional set of constraints. If the prior knowledge can be expressed as a multivariate Normal distribution with certain parameters, updating the background distribution remains tractable (timing experiments follow in Section III).

The SI of a pattern is defined as SI = IC/DL. The IC of a pattern Q is defined as the information gain, i.e., the increase in likelihood of the full data, from the current background distribution, to the background distribution also encompassing Q. Due to the context, this is equivalent to $-\log(\Pr(Q))$. That is, we have to compute the probability of the pattern being present under the background distribution.

The DL corresponds to the complexity of communicating the pattern to the user. Given that this depends on the number of conditions $|\mathcal{C}|$ used to describe the subgroup, the DL has the form $DL = \gamma|\mathcal{C}| + \eta\ (+1)$. Here, the $+1$ is for spread patterns, which have one more term than location patterns.

Computing the IC as well as updating the background distribution with patterns is a highly non-trivial exercise, but omitted here for brevity (see [3] for the full derivations).

TABLE I
CHANGE IN SI FOR THE TOP PATTERNS OVER FOUR ITERATIONS (§III).

Intention	SI Iter1	Iter 2	Iter 3	Iter 4
a3 = '1'	48.35	-1.13	-1.13	-1.13
a5 = '1'	47.49	47.49	-1.13	-1.13
a4 = '1'	39.49	39.49	39.49	-1.13
a3 = '1' ∧ a4 = '0'	36.26	-0.85	-0.85	-0.85
a3 = '1' ∧ a5 = '0'	36.26	-0.85	-0.85	-0.85
a5 = '1' ∧ a3 = '0'	35.62	35.62	-0.85	-0.85

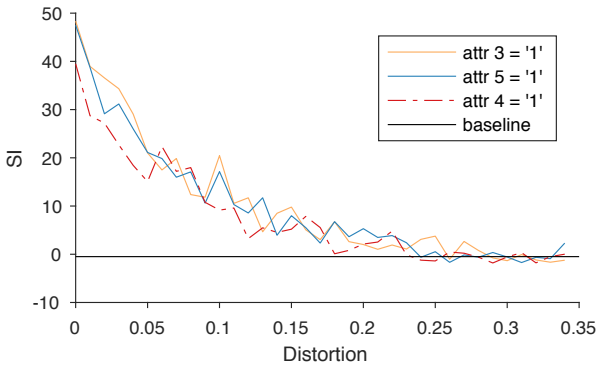

Fig. 3. SI of subgroups in the synthetic data, (§III), corresponding to true descriptions when adding and removing points randomly to the subgroups.

Search strategies. There appear to be no obvious efficient solutions to mine the best location or spread patterns. We resort to optimization procedures that are commonly used in either scenario. To find location patterns with high SI, we employ beam search. For spread patterns, we first search for the best location pattern and after updating the background distribution with the location, we use gradient descent and random restarts to find a weight vector $\mathbf{w}$ with high SI for that subgroup.

III. EXPERIMENTS

We evaluated whether our method is able to find good location and spread patterns in terms of SI and whether the model updates work as expected. We also studied the pattern descriptions, to see whether the patterns found appear to be interesting. We conducted experiments on four datasets: one synthetic and three publicly available ones, of varying nature. The results for each dataset are described in the following subsections. Finally, we studied the scalability of the methods. For details on the set-up and more extensive results see [3].

Synthetic data. We generated a dataset (shown in Fig. 2) with two target attributes and five binary description attributes. The first three description attributes contain the true labels for subgroups p_1 to p_3, the other two take random values.

We tested whether our method could reliably retrieve the embedded patterns. We corrupted the description attributes by randomly flipping every 0 and 1 with a certain probability. We performed the two-step spread pattern mining process for three iterations, and at each iteration we selected the top pattern to update the background distribution. Then, we checked up to what noise level the subgroups can still be retrieved.

We observed that our method correctly finds the embedded subgroups in the first three iterations (top pattern shown in Fig. 2b). It also retrieved the direction along which each subgroup's spread differs most from the full data covariance.

Table I shows the change in SI for the top six patterns from iteration one, in subsequent iterations. Once the embedded subgroups were selected and used to update the background distribution, the SI of those patterns and the SI of the derived patterns dropped and remained low afterwards. Hence, updating the background distribution and the influence that should have on the IC scores of patterns worked as expected.

It can be observed also that the subgroups with more complex descriptions (e.g., a3 = '1' ∧ a4 = '0') have lower SI. While their extension is equivalent to the corresponding a_i = '1' pattern, the SI is lower because their DL is higher. Non-redundancy in the description is indeed achieved naturally.

Fig. 3 gives the result of the retrieval experiment with noise added to the description attributes. We find that all embedded patterns can still be recovered when the flipping probability is up to 0.22, and partially retrieved up to 0.25. These values correspond to adding a random set of points that is roughly three and four times the size of the embedded pattern (e.g., $(1 - 0.25) \cdot 40 = 30$ vs. $0.25 \cdot 480 = 120$). We conclude that the method is quite robust against noise.

Socio-economics data. As a case study, we mined location and spread patterns in data encompassing voting percentages, age distributions, and workforce distribution statistics for the 2009 elections in Germany for all 412 administrative districts [4]. We used the vote count attributes as targets and the age and the work force attributes for the descriptions. Geolocations were used only for interpretation. To increase interpretability, we enforced a 2-sparsity constraint on the spread patterns.

Fig. 4a shows the top location pattern, and Fig. 4b,c some explanation and the top spread pattern. Comparing the distribution of the top pattern against the expected distribution under the model (Fig. 4b, red and blue lines), we observe that the voting behavior in the covered districts deviates substantially from the full population: more votes for Left, fewer for all others. The intention of the pattern corresponds to districts with relatively few children; from Fig. 4a we see the extension covers mainly East Germany.

Once we update the background distribution with the location pattern, the model mean for the subgroup becomes the observed mean (Fig. 4b). Then, we find that the spread pattern with highest SI is related to the covariance between the Social Democrats (SPD) and Christian Democrats (CDU/CSU), with weight vector $(0.5704, 0.8214$; Fig. 4c). The variance in this direction is much smaller than expected. Since the votes add up to a constant, we also expect negative correlations between the parties under the model, but for this subgroup the anti-correlation is much stronger than expected (these parties are in heavier competition here). In our opinion, these patterns appear to convey potentially highly interesting insights.

Scalability. We have not studied the algorithmic problem of how to find good solutions in depth. The computation time

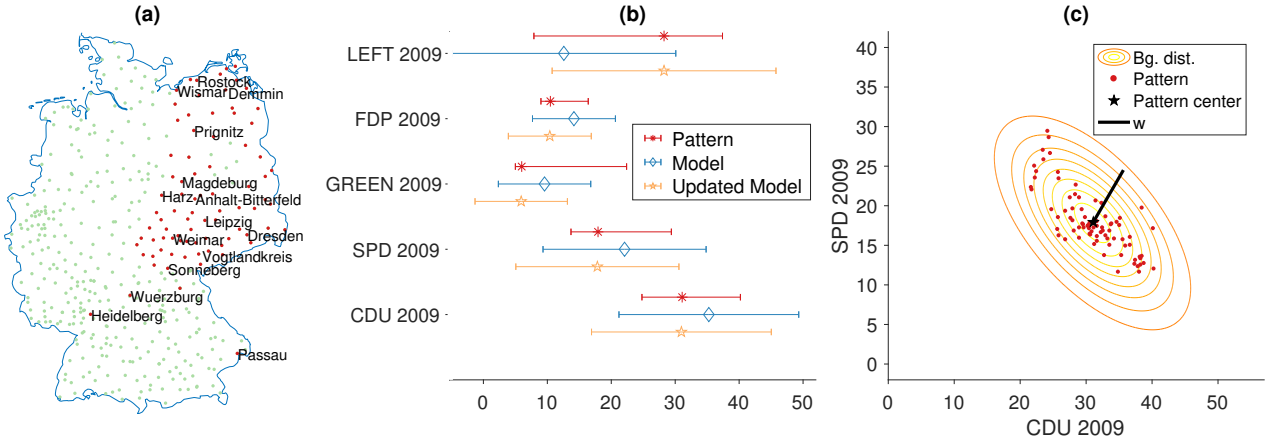

Fig. 4. Top subgroup pattern in the Socio-economics data (§III), "Children Pop. $<= 14.1$". (a) Districts covered by the subgroup, (b) comparison of vote distribution for the model and the covered districts, (c) top spread pattern.

TABLE II
RUNTIME TO UPDATE THE BACKGROUND DISTRIBUTION WITH IDENTIFIED PATTERNS. FIRST ROW SHOWS TIME (IN SECONDS) TO FIT THE INITIAL DISTRIBUTION, CONSECUTIVE ROWS TIME UNTIL CONVERGENCE WHEN INCORPORATING ADDITIONAL PATTERNS. DATA SETS: GERMAN SOCIO-ECONOMICS (GSE; $n = 412$, $d_{\mathbf{x}} = 13$, $d_{\mathbf{y}} = 5$), WATER QUALITY (WQ; $n = 1060$, $d_{\mathbf{x}} = 14$, $d_{\mathbf{y}} = 16$), CRIME (CR; $n = 1994$, $d_{\mathbf{x}} = 122$, $d_{\mathbf{y}} = 1$), MAMMALS (MA; $n = 2220$, $d_{\mathbf{x}} = 67$, $d_{\mathbf{y}} = 124$).

	Location pattern				Spread pattern		
Iteration	GSE	WQ	Cr	Ma	GSE	WQ	Cr
Init	9.167	8.640	9.714	8.453			
1	0.13	0.16	0.12	13.72	0.10	0.10	0.11
2	0.09	0.16	0.08	33.09	0.08	0.05	0.08
3	0.12	0.31	0.09	62.61	0.06	0.12	0.09
4	0.25	0.52	0.11	120.44	0.11	0.13	0.13
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	1.49	4.00	0.80	1130.81	0.42	0.46	0.83

of the beam search algorithm can be controlled through the search parameters, and it employs a timer. This strategy allows it to work on data of any size and dimensionality. Likewise, the heuristic solution to mine spread patterns typically outputs a pattern in very little time. Notice that for both algorithms, the runtime is linear in the number of data points.

The computational cost of fitting the background distribution is less obvious. Results for incorporating location and spread patterns into the background distribution are presented in Table II. We find that for the Mammals data, which has target dimension 124, the time quickly grows to durations that cannot be considered acceptable for interactive use. For spread patterns, this problem does not occur because they are by definition of low rank (it is a one-dimensional projection).

IV. CONCLUSION

Numerous unsupervised methods exist to make sense of real-valued datasets, most notably methods for dimensionality reduction and clustering. Labels (or description attributes as in this paper) associated with the data points are then often used to interpret these results, e.g., by measuring enrichment of certain labels within a cluster. However, whether that provides explanations or insights is a matter of coincidence: there is no a priori reason that clusters should be enriched, or a guarantee that equally colored points are grouped in a scatter plot.

We propose an alternative approach, by using the description attributes to guide the search for surprising multivariate relations in the data. Resulting subgroups are then automatically explained well by the descriptions. Our approach contrasts with traditional supervised methods in focusing on *local* patterns: properties of the target attributes that apply only to subsets of the data. Arguably, due to the increasing amount and the resulting inhomogeneity of data, the importance of local patterns is bound to increase.

Our approach generalizes the literature on Subgroup Discovery and Exceptional Model Mining in being applicable for real-valued target attributes of arbitrary dimensionality, and in searching for multivariate local patterns across all these dimensions, including unusual covariance structures. Moreover, the interestingness of the patterns of this type is formalized rigorously, quantifying the amount of information the user gains by observing them. We found that the resulting algorithms are effective and efficient, in theory and in practice.

Acknowledgements. This work has been supported by the ERC under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. 615517, FWO (project no. G091017N, G0F9816N), the European Union's Horizon 2020 research and innovation programme and the FWO under the Marie Skłodowska-Curie Grant Agreement no. 665501, the Academy of Finland (288814, 313513), and Tekes (Revolution of Knowledge Work project).

REFERENCES

[1] T. De Bie, "An information theoretic framework for data mining," in *Proc. of KDD*, 2011, pp. 564–572.
[2] M. Meeng and A. Knobbe, "Flexible enrichment with Cortana – software demo," in *Proc. of BeneLearn*, 2011, pp. 117–119.
[3] J. Lijffijt, B. Kang, W. Duivesteijn, K. Puolamäki, E. Oikarinen, and T. De Bie, "Subjectively interesting subgroup discovery on real-valued targets," arXiv:1710.04521 [stat.ML], 2017.
[4] M. Boley, M. Mampaey, B. Kang, P. Tokmakov, and S. Wrobel, "One click mining: Interactive local pattern discovery through implicit preference and performance learning," in *Proc. of KDD-IDEA Workshop*, 2013, pp. 27–35.