Subgroup Discovery through Bump Hunting on Relational Histograms

Radomír Černoch and Filip Železný

Czech Technical University in Prague, Faculty of Electrical Engineering,
Department of Cybernetics, Intelligent Data Analysis research lab

Abstract. We propose an approach to subgroup discovery in relational databases containing numerical attributes. The approach is based on detecting bumps in histograms constructed from substitution sets resulting from matching a first-order query against the input relational database. The approach is evaluated on seven data sets, discovering interpretable subgroups. The subgroups' rate of survival from the training split to the testing split varies among the experimental data sets, but at least on three of them it is very high.

1 Introduction

The goal of *subgroup discovery* [4] has been defined in [12] as "given a population of individuals and a property of individuals we are interested in, find population subgroups that are statistically most interesting, e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest." This viewpoint of subgroup discovery has been reflected both in the propositional [8,1] and relational [12,11] setting.

Here we explore a slightly different notion of subgroup discovery. To motivate it, consider first a single data table in which rows describe clients and one of the columns (attributes) contains the client's age, which is the property of interest. Since age is a numerical attribute, the most natural way to discover subgroups with respect to this property is to plot the histogram of age from the data table and look for possible bumps in the histogram. Analogically, bumps in multi-dimensional histograms may be considered when several numerical attributes are jointly of interest. Bump-hunting strategies for such tasks have been explored [3,13,9] in the propositional setting.

Here we adhere to the single-dimensional case but instead generalize the bump-hunting framework to the relational setting. For example, we would like to be able to discover in a relational banking database that clients from Prague fall apart into distinct subgroups according to the size of deposits they make. This can be accomplished by inspecting the histogram of the *Amount* variable's values in the set of substitutions making the query

$$\mathsf{client}(C) \wedge \mathsf{residence}(C, \mathsf{prague}) \wedge \mathsf{account}(A, C) \wedge \mathsf{deposit}(A, Amount)$$

true in the database. However, to preserve the interpretability of the histogram, there should be one element in the substitution set per each client. That is to say, values of *Amount* pertaining to a single client must be aggregated (e.g., averaged) in the substitution set before constructing the histogram.

Motivated by the above example, we developed an algorithm that, given a relational database with a distinguished main relation (such as client above), searches for a triple consisting of a query, a variable within the query, and a suitable aggregation function, so that the histogram of numerical values constructed from these three ingredients, in the way exemplified above, exhibits remarkable bumps. To this end, we address both the logical and statistical aspect of the algorithm: in particular, we design a refinement-based search for the target query as well as a fast, “visually inspired” histogram-inspection technique. The simplicity and speed of the latter is vital due to the generally daunting embedding of one data mining task (bump hunting) in another (query search).

In relational machine learning, the idea of analyzing substitution sets has been explored before, e.g., for constituting relational kernels [7] or for propositionalization [5]. However, we are not aware of a previous attempt to analyze histograms derived from substitution sets. Our approach also differs from current subgroup discovery systems, mainly in two respects. First, it does not assume a prior indication of a target property of interest. Second, it does not produce any explicit rules predicting memberships in the respective discovered subgroups. Indeed, note that the earlier example-query just specifies a database view from which a histogram is derived; the query thus defines the population rather than the subgroups. For these reasons, our approach falls fully in the unsupervised learning family, unlike the straddled systems RSD [11] or CN2-SD [8].

2 Algorithm

We first describe the inner component of the algorithm (bump hunting) and then the outer one (query search). In this short account, we omit pseudocodes.

2.1 Bump hunting

Briefly, we understand bumps as modes in histograms which are mutually separated by areas of low probability. A histogram is represented by the function $P(w)$, $w \in W$ where $W = \{1, 2, \ldots\}$ is a finite set of bin indexes. We extract the ordered subset $\{k_1, k_2, \ldots\} \in W$ of modes (local maxima) of $P(w)$. Then for its each two successive elements k_i, k_{i+1} we compute the (normalized) area of low probability

$$\mathsf{ALP}(i, W) = \frac{1}{|W|} \sum_{k_i < w < k_{i+1}} \min\{P(k_i), P(k_{i+1})\} - P(w) \qquad (1)$$

which is illustrated in Fig. 1. ALP grows with the size of the lateral modes and decays with the interleaving bin area. Informally, it grows both with the coverage and mutual separation of the subgroups. The heuristic is however not fully satisfactory as it would indicate a large number of subgroups in a histogram with

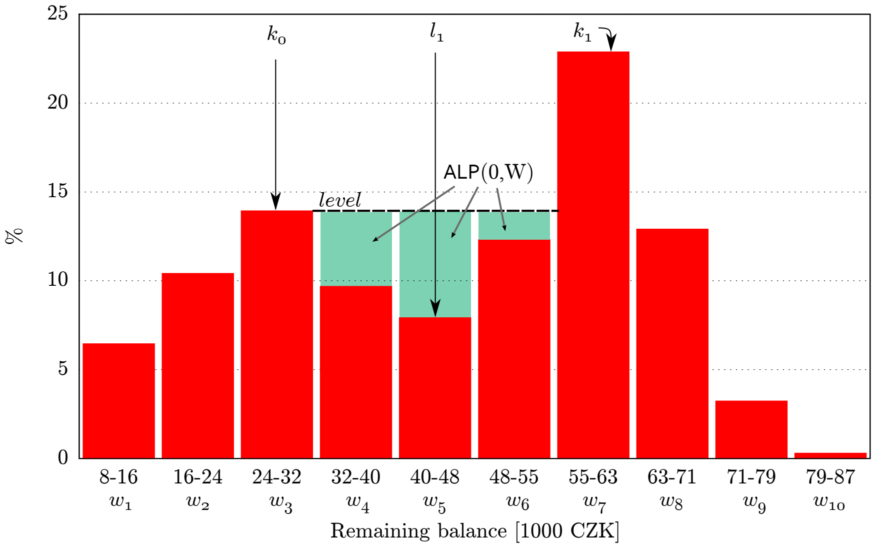

Fig. 1. An illustration of the ALP (area of low probability) concept on the financial dataset [2]. The random variable shown in the histogram is the highest balance ever to appear on client's account. Since the ALP is large here, the modes k_0 and k_1 are deemed subgroups.

a large number of local maxima and with roughly equal sizes of the ALP values. Such subgroups could be artifacts of noise, and even if not, they would not be very interesting. Therefore we use a 'noise cancelling' heuristic ncALP prioritizing histograms where only one of the areas of low probability is dominant.

$$\mathsf{ncALP}(W) = \max_{1\leq j\leq m} \frac{\mathsf{ALP}(j,W) - \mathsf{ALP}(j+1,W)}{j}$$

More exhaustive or statistically-principled approaches to bump identification have been proposed [3,13,9] using techniques such as genetic algorithms or the EM algorithm. Unlike these approaches, the above simple method clearly has a linear time complexity. Low runtime is vital here due to the embedding of bump-hunting inside the query-search.

2.2 Query search

We follow a beam-search strategy with refinement steps (adding an atom with fresh variables, unifying or instantiating a variable) according to a pre-defined language bias as usual in inductive logic programming. We start with an empty query. Unlike in standard ILP algorithms, each query in the search agenda is in exactly one of three successive phases.

Germination: before a query contains an atom with a numerical variable it can not generate any histogram and is not evaluated against the database.

Production: query contains one or more numerical variables and it generates histograms according to which it is evaluated. Multiple histograms are created, one for each combination of i) a numerical variable in the query, ii) an aggregation operator (drawn from a pre-defined set), and iii) the number of bins from a pre-defined interval. Out of these histograms, the one which maximizes the ncALP criterion is associated with the query modulo the following adjustment. To prevent the beam search from repeatedly finding certain subgroups, the ncALP heuristic value is multiplied by the distance (dissimilarity metric) between the evaluated histogram and all previously found.

Retirement: when a query becomes overly specific, the set of instantiations for a variable becomes empty. Such a query is discarded from the search agenda.

3 Experimental Evaluation

We tested the proposed algorithm on seven real-world datasets to see if it is able to discover non-trivial yet interpretable subgroups, and to quantify any possible tendency towards overfitting. We used the financial dataset [2], two parts of the mutagenesis dataset [10] (regression-friendly, -unfriendly), and four genomic datasets [6], each describing the relational structure of a biochemical pathway along with numerical values of expressions of the involved genes in a single phenotype (glioma samples).

Two examples of histograms with interpretable subgroups discovered in the financial database are shown in Fig. 2. The first example is interpretable given the regional division of the Czech Republic; unlike all other parts of the country, Prague constitutes a single region with urbanization reaching 100%. The bump at the right-hand side of the plot thus corresponds to loans awarded in Prague. The second example roughly shows there are small loans and big loans. The algorithm discovered the mentioned fact by querying for balances after transactions corresponding to a given loan, taking the maximum of these balances, and bump-hunting on the histogram of such obtained maxima.

Overfitting tendency was assessed using $k \times 2$ ($k = 6$) cross-validation in which the training and testing sets are equally large; this is needed for properly comparing the training and testing histograms. For each training split, 5 discovered triples (query, variable, aggregator) with the highest ncALP value were kept, and their ncALP′ values were computed on the corresponding test split. As a result, we obtained $2 * 6 * 5 = 60$ ratio values $r = \mathsf{ncALP}'/\mathsf{ncALP}$. A single representative value, called subgroup stability, is then computed as the geometric mean $R = \sqrt[60]{r_1 \cdot \ldots \cdot r_{60}}$.[1] This value decreases with overfitting: Ideally $R = 1$, but if bumps found in the training set do not exists in the testing set, then R goes to 0. Note that if there is a single value of $r = 0$, the value of R becomes 0 destroying any information in other r values. To overcome this problem, the amount of zero values among the r's was saved as D and then R was computed only from positive values of r. E.g. $D = 20\%, R = 75\%$ means that one fifth of histograms in the testing set contain no bump and the remaining ones are 1/4 less significant on average. Table 3 shows the results for the seven data sets.

[1] The geometric mean is better suited for ratios than the widely used arithmetic mean. E.g. mean value of 10% and 1000% is by common sense 100% rather than 505%.

Dataset	Disappeared bumps (D [%])	Subgroup stability (R [%])	Computation time ([s])
Mutagenesis - easy	0	90	311
Mutagenesis - hard	0	64	153
Financial	4	48	1871
hsa00190-genes expr. in glioma	30	41	444
hsa04360-genes expr. in glioma	26	37	363
hsa01430-genes expr. in glioma	20	32	335
hsa04514-genes expr. in glioma	37	15	3270

Table 1. Overfitting tendency of the subgroup discovery algorithm expressed through two quantities (see main text).

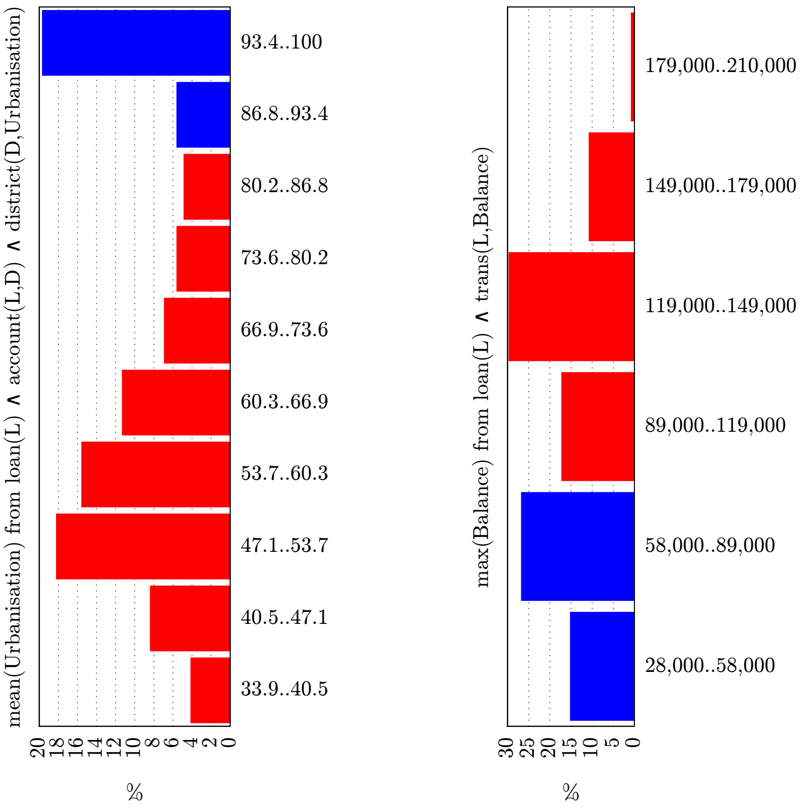

Fig. 2. Examples of discovered subgroups. Left: Mean of relative urbanization (percentage of population living in cities) of regions, where clients live. Right: Max balance after executing a transaction. In both cases, the values of the analyzed variable (A) in the substitution set are aggregated (left: mean, right: max) with respect to the main relation loan so that the vertical axis corresponds to the frequency of loans.

4 Conclusions

We proposed a new concept of relational subgroup discovery based on searching bumps in histograms constructed from substitution sets derived from matching a first-order query against a database. The fact that a relatively simple implementation of the concept was able to discover non-trivial interpretable subgroups that generally survive from training data to testing data is reassuring and calls for further advancement of the algorithm. This should mainly focus on improving the search strategy. Currently the search for the query, the optimization of the number of bins, and the evaluation of the histogram-heuristic are rather loosely coupled. A tighter integration would likely improve the overall performance, and might allow to find more complex patterns.

References

1. Atzmueller, M., Lemmerich, F.: Fast subgroup discovery for continuous target concepts. In: Foundations of Intelligent Systems. Springer (2009)
2. Berka, P., Sochorová, M.: Guide to the financial data set (1999), `http://lisp.vse.cz/pkdd99/berka.htm`
3. Friedman, J.H., Fisher, N.I.: Bump hunting in high-dimensional data. Statistics and Computing 9, 123143 (1999)
4. Kralj-Novak, P., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research 10, 377–403 (2009)
5. Krogel, M.A., Wrobel, S.: Transformation-based learning using multirelational aggregation. In: Inductive Logic Programming, pp. 142–155. Springer (2001)
6. Kuželka, O., Szabóová, A., Holec, M., Železný, F.: Gaussian logic for predictive classification. In: ECML/PKDD'2011: Eur. Conf. on Machine Learning / Principles and Practice of Knowledge Discovery in Databases. Springer (2011)
7. Landwehr, N., Passerini, A., De Raedt, L., Frasconi, P.: Fast learning of relational kernels. Machine Learning 78(3), 305–42 (2010)
8. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188 (2004)
9. Lowthian, P., Thompson, M.: Bump-hunting for the proficiency tester – searching for multimodality. The Analyst 127(10), 1359–1364 (2002)
10. Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E., King, R.D.: Theories for mutagenicity: A study in first-order and feature-based induction. Artificial Intelligence 85, 277–299 (1996)
11. Železný, F., Lavrač, N.: Propositionalization-based relational subgroup discovery with RSD. Machine Learning 62(1-2), 33–63 (2006)
12. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Principles of Data Mining and Knowledge Discovery, pp. 78–87. Springer (1997)
13. Yukizane, T., Ohi, S.Y., Miyano, E., Hirose, H.: The bump hunting method using the genetic algorithm with the extreme-value statistics. IEICE - Trans. Inf. Syst. E89-D, 2332–2339 (2006)