

Subgroup Discovery Method SUBARP

Klaus Truemper

Department of Computer Science,
University of Texas at Dallas,
Richardson, TX 75083, U.S.A.
truemper@utdallas.edu

Abstract. This paper summarizes the subgroup discovery method SUBARP. The discussion is to provide an intuitive understanding of the ideas underlying the method, and mathematical details are omitted.

Key words: subgroup discovery, logic formula, explanation

1 Introduction

Suppose we have a collection of *records*, each of which contains values for a set of *attributes*. The attribute values may be nominal, for example, "red," "green," "yellow"; or may be discrete, for example, "1", "2", "3"; or may be continuous, for example, "3.174", "5.091", "20,100.56." Some values in a record may not be known, indicated by the entry "?." *Subgroup Discovery* aims at finding important relationships that are implicitly contained in the data; see [5, 13].

There are a number of Subgroup Discovery methods; for a survey that unifies Subgroup Discovery, *Contrast Set Mining*, and *Emerging Pattern Mining*, see [7]. In particular, any standard classification rule learning approach can be adapted to Subgroup Discovery [8]. SUBARP is one such method, adapted from the classification method Lsquare [1–4]. The overview of SUBARP below is quite complete, except that rather technical discretization aspects are ignored; for details, see [1, 9, 10]. It is assumed that all nominal entries of records have been converted to integers prior to application of SUBARP.

2 Subgroup Discovery Method SUBARP

As part of the input, SUBARP requires that at least one attribute is defined to be a *target*. The method processes each target separately and attempts to find relationships that explain the target values by the values of remaining attributes. From these relationships, the subgroups are determined.

3 Target Discretization

Suppose there is an interesting subgroup where the values of a target t fall into a certain range R , say $R = \{c \leq t \leq d\}$. If we knew c and d , then we could

search for the subgroup by comparing the records where t falls into the range R with the records where t is outside R . Absent prior knowledge of such c and d , SUBARP computes potentially useful target ranges R ; an auxiliary tool for this task is described in [9].

For given target range R , let A be the set of given records whose target values t are in R , and define B to be the set of remaining records. SUBARP tries to explain the difference between the sets A and B in a multi-step process. The steps involve feature selection, computation of explanations, factor selection, subgroup evaluation, and tests of statistical significance. They are described in subsequent sections. The first two steps are implemented using Lsquare cited above.

We need to introduce the notion of a *logic formula* that to a given record assigns the value *True* or *False*. The concept is best explained by an example. Suppose that a record has x and y among its attributes, and that the record values for these attributes are $x = 3.5$ and $y = 4.7$. The logic formula consists of *terms* that are linear inequalities involving the attributes, and the operators \wedge ("and") and \vee ("or"). For a given record, a term evaluates to *True* if the inequality is satisfied by the attribute values of the record, and evaluates to *False* otherwise. Given *True/False* values for the terms, the formula is evaluated according the standard rules of propositional logic.

For example, $(x \leq 4)$ and $(y > 3)$ are terms. A short formula using these terms is $(x \leq 4) \wedge (y > 3)$. For the above record with $x = 3.5$ and $y = 4.7$, both terms of the formula evaluate to *True*, and hence the formula has that value as well. As a second example, $(x \leq 4) \vee (y > 5)$ produces the value *True* for the cited record, since at least one of the terms evaluates to *True*.

4 Feature Selection

This step, which is almost entirely handled by Lsquare, repeatedly partitions the set A into subsets A_1 and A_2 ; correspondingly divides the set B into subsets B_1 and B_2 ; finds a logic formula that achieves the value *True* on the records of A_1 and *False* on those of B_1 ; and tests how often the formula achieves *True* for the records A_2 and *False* for those of B_2 . In total, 40 formulas are created. The frequency with which a given attribute is used in the 40 formulas is an indicator of the importance of the attribute in explaining the differences between the sets A and B . Using that indicator, a significance value is computed for each attribute. The attributes with significance value beyond a certain threshold are selected for the next step.

5 Computation of Explanations

Using the selected attributes, Lsquare computes two formulas. One of the formulas evaluates to *True* for the records of A and to *False* for those of B , while the second formula reverses the roles of A and B .

Both formulas are in *disjunctive normal form* (DNF), which means that they consist of one or more *clauses* combined by "or." In turn, each clause consists of linear inequality terms as defined before and combined by "and." An example of a DNF formula is $((x \leq 4) \wedge (y > 3)) \vee ((x < 3) \wedge (y \geq 2))$, with the two clauses $((x \leq 4) \wedge (y > 3))$ and $((x < 3) \wedge (y \geq 2))$. Later, we refer to each clause as a *factor*.

We discuss below how subgroups are derived from the first DNF formula. The same process is carried out for the second DNF formula, except for reversal of the roles of A and B .

6 Factor Selection

Recall that the first DNF formula evaluates to *False* for the records of B . Let f be a factor of that formula. Since the formula consists of clauses combined by "or," the factor f also evaluates to *False* for all records of B . Let S be the subset of records for which f evaluates to *True*. Evidently, S is a subset of A .

We introduce a direct description of S by an example. Suppose the factor f has the form $f = ((x \leq 4) \wedge (y > 3))$. Let us assume that A was defined as the subset of records for which the values of target t fall into the range $R = \{2 \leq t \leq 9\}$. Then S consists of the records of the original data set satisfying $2 \leq t \leq 9$, $x \leq 4$, and $y > 3$. But we know more than just this characterization. Indeed, each record of B violates at least one of the inequalities $x \leq 4$ and $y > 3$. Put differently, for any record where t does not satisfy $2 \leq t \leq 9$, we know that at least one of the inequalities $x \leq 4$ and $y > 3$ is violated.

So far, we have assumed that the formula evaluates to *False* for all records of B . In general, this goal may not be achieved, and the formula produces the desired *False* values for most but not all records of B . Put differently, S then contains a few records of B . For the above example, this means that for some records where the target t does not satisfy $2 \leq t \leq 9$, both inequalities $x \leq 4$ and $y > 3$ may actually be satisfied.

Potentially, the target range R and the factor f characterize an important configuration that is interesting and useful for experts. On the other hand, the information may be trivial and of no interest. To estimate which case applies, SUBARP computes a significance value for each factor that lies in the interval $[0, 1]$. The rule for computing the significance value is similar to those used in other Subgroup Discovery methods, where the goal of *interestingness* [6] is measured with quality functions balancing conflicting goals such as (1) the size of the subgroup, (2) the length of the pattern or rule defining the subgroup, and (3) the extent to which the subgroup is contained in the set A . Specifically, SUBARP decides significance using the third measure and the probability with which a certain random process creates a set of records T for which the target t is in R , the factor f evaluates to *True*, and the size of T matches that of S . We call that random process an *alternate random process* (ARP).

Generally, an ARP is an imagined random process that can produce an intermediate or final result of SUBARP by random selection. SUBARP considers

and evaluates several ARPs within and outside the Lsquare subroutine, and then structures decisions so that results claimed to be important can only be achieved by the ARPs with very low probability. The overall effect of this approach is that subgroups based on factors with very high significance often turn out to be interesting and important when evaluated by experts.

The next step evaluates subgroups.

7 Evaluation of Subgroups

When a data set is to be analyzed, it first is split randomly, usually with 50/50 ratio, into a training set and a testing set. In another approach, preferred by us, the data set is first sorted according to the target values. Consider the sorted records to be numbered $1, 2, \dots$. Then all odd numbered records are placed into the training set and all even numbered ones into the testing set. By this approach, the entire range of target values is properly represented in both files.

Regardless of the way the training and testing sets are selected, the testing set is put aside, and only the training set is used by SUBARP to derive subgroups and their significance values as described above. Once the significance values have been computed for the subgroups, the associated target ranges and factors are used to compute a second significance value using the testing data. The average of the significance values obtained from the training and testing data is assigned as *overall significance* value to each subgroup.

In numerous tests of data sets, it has turned out that only subgroups with very high significance value are potentially of interest to experts and thus may produce new insight. Based on these results, only subgroups with significance above 0.90 are considered potentially useful. In exceptional case, this threshold may be lowered, but never below 0.80. Even when a large number of subgroup candidates have been originally computed, the threshold typically leaves very few surviving subgroups. These subgroups are subjected to tests of statistical significance that constitute another hurdle before final acceptance.

8 Tests of Statistical Significance

As before, let R be the range and f be the factor of a given subgroup S derived from the sets A and B . By the derivation of S , we know that, for each training record for which f evaluates to *True*, the target value t very likely is in the range R .

For statistical verification that this fact holds generally for records of the underlying population, we consider an alternative hypothesis that there is no relationship between *True* evaluations of f and membership of t in R . Specifically, the alternative hypothesis is that, during random selection of a record, it is randomly decided with some probability p whether f of the record evaluates to *True*, regardless of whether the target value of the record falls into the range R or not.

Let C be the subset of the testing records for which the target value t is outside R . Define n to be the number of records of C . Suppose that, for k records of C , the factor f evaluates to *True*. The probability that we see at most k records of C where f evaluates to *True* under the hypothesis, is $Q = \sum_{j=0}^{j=k} \binom{n}{j} p^j (1-p)^{n-j}$. For the evaluation of Q , we estimate p from the testing data. If Q is very small, then the hypothesis unlikely holds and is rejected. In that case, we accept the subgroup as being important. The numerical value associated with the criterion "very small" depends on the application. For example, in medicine, it often means 0.0001, while in other settings, values up to 0.01 may be acceptable for the rejection test.

The user may apply other statistical tests. We sketch an example using the above notation. Let q be the ratio of k divided by the total number of testing records for which f evaluates to *True*. *A priori* we defined null and alternate hypotheses postulating certain bounds for q . Evaluation of the two hypotheses via the binomial distribution decides which hypothesis is more likely correct.

9 Multiple Targets

The description so far covers processing of a single target.

In the general case, several targets are of interest. A difficulty arises when a target is similar to another target. In that setting, SUBARP may explain one target in terms of a related one. To prevent this useless result, equivalence groups of attributes can be defined where a target falling into a group may not be explained using any other attribute in that group. The equivalence groups may intersect, with no adverse effect on the computations.

10 Multiple Explanations

It is possible that several explanations of a target are possible, and each of them is of interest to the user. In the computations as described above, one explanation may mask other explanations. This negative effect is avoided as follows.

When a number of explanations have been found, the most significant one is analyzed, and one attribute of that explanation is made inaccessible to SUBARP. Then the entire process is started again. As a result, the leading explanation cannot be reproduced, and other explanations may surface. The process stops when no additional explanation can be found.

The process may produce a lot of explanations. Indeed, many of the explanations may be identical, or some explanation may dominate others. SUBARP weeds out duplicate or dominated explanations, and outputs the surviving cases.

11 Complexity of Explanations

Reference [11] describes an extension of Lsquare where, in the present terminology, the inequalities defining factors are not restricted to one variable each.

The extension allows control over the number of variables used in such inequalities. In experiments, an upper limit of 2–4 has turned out to be useful, see [11]. When that extension is invoked in SUBARP, the factors of subgroups have the more general format. This feature supports detection of important subgroups that cannot be found when each inequality is restricted to just one variable. Geometrically speaking, the standard Lsquare method leads to subgroups that correspond to rectangular boxes, while the extension produces subgroups represented by more general polyhedra.

Reference [11] also investigates the comprehensibility of explanations. Generally speaking, SUBARP is so constrained that the explanations are comprehensible by humans. This allows intuitive validation by experts and greatly contributes to acceptance of explanations.

Human comprehensibility is somewhat related to the concept of VC dimension [12]. In the present setting, low values of the VC dimension are attractive. SUBARP computes an estimate of the VC dimension to help with the evaluation of the comprehensibility of explanations. For details, see [11].

12 Challenging Cases

Subgroup Discovery is particularly challenging when data sets have few records with large number of attributes, since then for a given target range numerous factors may exist. Such a setting occurs frequently in medical research; or when the assembly of records is costly or time-consuming; or when there simply are not many records, yet the number of attributes that possibly explain targets is large or even huge.

SUBARP handles such cases with ease. For example, it has successfully processed record sets with less than 40 records and more than 400 attributes.

References

1. S. Bartnikowski, M. Granberry, J. Mugan, and K. Truemper. Transformation of rational and set data to logic data. In *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*. Springer, 2006.
2. G. Felici, F. Sun, and K. Truemper. Learning logic formulas and related error distributions. In *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*. Springer, 2006.
3. G. Felici and K. Truemper. A MINSAT approach for learning in logic domain. *INFORMS Journal of Computing*, 14:20–36, 2002.
4. G. Felici and K. Truemper. The lsquare system for mining logic data. In *Encyclopedia of Data Warehousing and Mining*, pages 693–697. Idea Group Publishing, 2005.
5. W. Klösgen. EXPLORA: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.
6. W. Klösgen. Subgroup discovery. In *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, 2002.

7. P. Kralj Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.
8. N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning*, 57:115–143, 2004.
9. K. Moreland and K. Truemper. Discretization of target attributes for subgroup discovery. In *Proceedings of International Conference on Machine Learning and Data Mining (MLDM 2009)*, 2009.
10. J. Mugan and K. Truemper. Discretization of rational data. In *Proceedings of MML2004 (Mathematical Methods for Learning)*. IGI Publishing Group, 2007.
11. K. Truemper. Improved comprehensibility and reliability of explanations via restricted halfspace discretization. In *Proceedings of International Conference on Machine Learning and Data Mining (MLDM 2009)*, 2009.
12. V. Vapnik. *The Nature of Statistical Learning Theory (Second edition)*. Springer-Verlag, 2000.
13. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of First European Conference on Principles of Data Mining and Knowledge Discovery*, 1997.