The Computer Journal Advance Access published September 22, 2012

doi:10.1093/comjnl/bxs132

Contrasting Subgroup Discovery

Laura Langohr[1]*, Vid Podpečan[2,3], Marko Petek[4], Igor Mozetič[2], Kristina Gruden[4], Nada Lavrač[2] and Hannu Toivonen[1]

[1]*Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland*
[2]*Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia*
[3]*International Postgraduate School Jožef Stefan, Ljubljana, Slovenia*
[4]*Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*
**Corresponding author: laura.langohr@cs.helsinki.fi*

Subgroup discovery methods find interesting subsets of objects of a given class. Motivated by an application in bioinformatics, we first define a generalized subgroup discovery problem. In this setting, a subgroup is interesting if its members are characteristic for their class, even if the classes are not identical. Then we further refine this setting for the case where subsets of objects, for example, subsets of objects that represent different time points or different phenotypes, are contrasted. We show that this allows finding subgroups of objects that could not be found with classical subgroup discovery. To find such subgroups, we propose an approach that consists of two subgroup discovery steps and an intermediate, contrast set definition step. This approach is applicable in various application areas. An example is biology, where interesting subgroups of genes are searched by using gene expression data. We address the problem of finding enriched gene sets that are specific for virus-infected samples for a specific time point or a specific phenotype. We report on experimental results on a time series dataset for virus-infected *Solanum tuberosum* (potato) plants. The results on *S. tuberosum*'s response to virus-infection revealed new research hypotheses for plant biologists.

Keywords: subgroup discovery; gene set enrichment

Received 30 November 2011; revised 15 August 2012
Handling editor: Einoshin Suzuki

1. INTRODUCTION

Subgroup discovery [1, 2] is a typical task in data mining for finding interesting subsets of objects. Classical subgroup discovery methods consider a set of objects interesting if they share a combination of attribute values that is characteristic for some class. In contrast, we aim to find subgroups of the following type: a set of objects is interesting if each of its members is characteristic for its own class, even if the classes are not identical. This allows finding patterns that could not be found with classical subgroup discovery.

For instance, in a dataset of bank customers, it may be the case that males tend to be characteristic in the sense that the combination of their education, occupation and location is characteristic for either high or low spenders. The setting proposed in this paper allows discovering males as an interesting subgroup, since being male implies that the person is characteristic for his class. Classical subgroup discovery methods would only be able to find separate subgroups for high spenders and low spenders, and would miss that males, in general, are characteristic for their classes.

This powerful effect is obtained by allowing the user to specify subsets of objects she wants to contrast in a flexible manner. First, these contrast sets can be defined using not only the original attributes, but also using information about characteristics with respect to classes (i.e. classical subgroup memberships). Secondly, contrast set definitions can use set-theoretic operations. For instance, an economist might be interested in contrasting different time points (e.g. before, during and after the financial crisis). She could then specify that she is interested in objects at a specific time point in contrast to all other time points. In such settings, classical subgroup discovery can contrast two time points, or several time points in a pairwise fashion. In the setting proposed here, and in the biological application that motivates our work, we are interested in contrasting subgroups from several time points (or several phenotypes) at the same time. We call this generalized problem *the contrasting subgroup discovery problem*.

To find such generalized subgroups of objects, we propose an approach that consists of two subgroup discovery steps and an intermediate, contrast set definition step. In the first step, interesting subgroups are found in a classical manner, based on semantic and statistical properties of objects. In the banking example, we can use an existing subgroup discovery method to find classical subgroups for the classes of low and high spenders, and would do this for each time point separately. In the second step, the user defines two new classes of objects; these are the contrast classes for the third step. As mentioned above, the definitions of contrast classes can take into account several different class attributes (such as different time points) as well as subgroup memberships from the first step. In the third and final steps, a classical subgroup discovery method is used to find interesting subgroups of objects of the two contrast classes. As a result, the subgroups can contain objects that are characteristic for their class, regardless of their class.

In the next section, we give a brief overview of classical subgroup discovery and describe how subgroup discovery and contrast mining have been addressed in different applications before (Section 2). In Section 3, we then propose the problem of contrasting subgroup discovery more formally. We then show how well-known algorithms can be combined to solve the problem as outlined above (Section 4).

In the second half of the paper, we focus on an important application in biology. In Section 5, we describe a gene set enrichment problem where the goal is to analyze contrasting gene sets, and we give an instance of the proposed methodology to solve the problem. In Section 6, we apply it on a time-series dataset from virus-infected potato plants (*Solanum tuberosum*) and report experimental results. Finally, we conclude with some notes about the results and future work.

2. BACKGROUND

Discovering patterns in data is a classical problem in data mining and machine learning [3, 4]. To represent patterns in an explanatory form, they are often described by rules $X \mapsto Y$, where $\mapsto$ denotes an implication and the antecedent X and the consequent Y can represent sets of attribute values (e.g. terms), a class or sets of objects, depending on the problem at hand.

Next, we define the problem of subgroup discovery formally, review related work and discuss how our approach differs from other pattern discovery approaches.

2.1. Subgroup discovery

Subgroup discovery methods find rules of the form $Condition \mapsto Subgroup$, where the antecedent $Condition$ is a conjunction of attribute values and the consequent $Subgroup$ is a set of objects, which satisfy some class-related interestingness measure.

Subgroups defined by individual attribute values. Consider a set S of objects, annotated by a set T of attribute values (e.g. terms). Each attribute value $t \in T$ defines a subgroup $S_t \subset S$ that consists of all objects $s \in S$ where t is true, that is, all objects annotated by the attribute value t:

$$S_t = \{s \in S \mid s \text{ is annotated by } t\}. \tag{1}$$

Example 2.1. Consider the bank customers of Table 1 which are annotated by the attributes *Occupation* and *Location* and assigned the class *high* or *low* for the class attribute *Spending* for two different time points, before and after the financial crisis, respectively. The attribute value $Location = village$ defines the subgroup $\{19, 20\}$ of two bank customers and $Occupation = education$ defines a subgroup of five bank customers $\{6, 9, 11, 16, 18\}$.

Subgroups defined by logical conjunctions of attribute values. Subgroups can be constructed by intersections, which are described by logical conjunctions of attribute values. Let $S_1, \ldots, S_k$ be k subgroups described by the attribute values $t_1, \ldots, t_k$. Then the logical conjunction of k attribute values defines the intersection of k subgroups:

$$t_1 \wedge t_2 \wedge \cdots \wedge t_k \mapsto S_1 \cap S_2 \cap \cdots \cap S_k. \tag{2}$$

TABLE 1. Bank customers *before* and *after* the financial crisis described by attributes Occupation and Location, and the class attribute Spending (adapted from [43]).

			Spending	
ID	Occupation	Location	Before	After
1	Industry	Big city	High	High
2	Industry	Big city	High	Low
3	Retail	Big city	High	Low
4	Finance	Big city	High	High
5	Doctor	Big city	High	High
6	Education	Big city	High	Low
7	Nurse	Big city	High	Low
8	Industry	Small city	High	High
9	Education	Small city	High	Low
10	Retail	Small city	High	Low
11	Education	Big city	Low	Low
12	Nurse	Big city	Low	Low
13	Unemployed	Big city	Low	Low
14	Retail	Small city	Low	Low
15	Doctor	Small city	Low	Low
16	Education	Small city	Low	Low
17	Unemployed	Small city	Low	Low
18	Education	Small city	Low	Low
19	Unemployed	Village	Low	Low
20	Unemployed	Village	Low	Low

Alternatively, we can write $T' \mapsto S_{T'}$, where T' is a set of attribute values $T' = \{t_1, \dots, t_k\} \subset T$, whose conjunction defines the subgroup $S_{T'} = S_1 \cap S_2 \cap \dots \cap S_k$.

Example 2.2. The set $T' = \{education, small\ city\}$ defines a subgroup of three bank customers {9, 16, 18} in Table 1.

An object can be a member of several subgroups. A subgroup might be a subset of another subgroup. In particular, in case the attribute values are organized in a hierarchy (or ontology), an object that is annotated by the attribute value t is also considered to be annotated by the ancestors of t in the hierarchy.

Example 2.3. Consider the hierarchies in Fig. 1. All individuals working in the retail sector also work in the service and private sector.

An *ontology* is a representation of a conceptualization and is often represented by a hierarchy, where nodes represent concepts (e.g. occupations or locations) and edges a subsumption relation (e.g. 'is a' or 'part of') between concepts [5]. See, for example, Fig. 1, where *nurses* as well as *doctors* are part of the *health* sector, which is part of the *public* sector. Ontologies can be used to incorporate background knowledge about attribute values (such as concepts, terms or something else). Subgroup discovery methods often use hierarchies to restrict the search space (see, e.g. [6, 7]), but subgroup discovery does not require that the attribute values be organized in a hierarchy.

Class-related interestingness measure. For each subgroup, one has to measure whether the subgroup is interesting or not. Classical subgroup discovery methods look for groups that are specific for a class when compared with the rest of the objects.

Similarly, to attribute values, classes define subgroups (sets) of objects. Let $c \in T$ be a specific class. Then an object $s \in S$ belongs to the subgroup defined by c if and only if s is annotated by c.

Example 2.4. Consider again the bank customers in Table 1. For the time point before, the financial crisis the class $Spending = high$ defines the subgroup $\{1, \dots, 10\}$ and the class $Spending = low$ defines the subgroup $\{11, \dots, 20\}$.

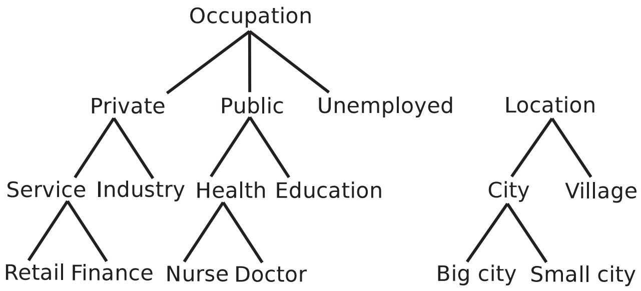

FIGURE 1. Example hierarchies of attribute values which are in this case terms (adapted from [43]).

In practice, a subgroup is often classified homogeneously. To formalize this idea, let

$$classes: \mathcal{P}(S) \rightarrow \mathbb{Z}_+ \times \mathbb{Z}_+ \tag{3}$$

be a function that gives the class distribution of a given set $S_{T'} \subset S$ of objects, that is, the number of objects in $S_{T'}$ annotated by c and the number of objects in $S_{T'}$ not annotated by c. (Here $\mathcal{P}(S)$ is the powerset of S.)

Example 2.5. Consider the subgroup $S_{T'} = \{9, 16, 18\}$ of three bank customers described by $T' = \{education, small\ city\}$ and the class $Spending = high$ for the time point before the financial crisis in Table 1. The class distribution of $S_{T'}$ is $classes(S_{T'}) = (1, 2)$ as one object of $S_{T'}$ is annotated by $Spending = high$ and two objects by $Spending = low$. Similarly, the class distribution of $S \setminus S_{T'}$ is $classes(S \setminus S_{T'}) = (9, 8)$.

Definition 2.1. *The classical* class-related interestingness measure *is a function*

$$\begin{aligned} f_c &: \mathcal{P}(T) \rightarrow \mathbb{R}, \\ &T' \mapsto g(classes(S_{T'}), classes(S \setminus S_{T'})), \end{aligned} \tag{4}$$

for some function g. That is, f_c is a function $g(\cdot)$ of the class distributions within and outside of the subgroup.

The exact definition of g varies from one problem variant to another, but the common denominator is that it is based on the class distributions alone.

Often the subgroups are analyzed by statistical tests, like Fisher's exact test, χ^2 test or the binomial probability. In our experiments, we use a p-value estimate obtained by *Fisher's exact test* [8] and a simple permutation test as class-related interestingness measure f_c. Without loss of generality, in the rest of the paper we assume that smaller values of f_c indicate more interesting subgroups.

Given a class attribute c with two possible classes $t_c, t_{\bar{c}} \in T$, the data are arranged in a contingency table for each subgroup $S_{T'} \subset S$, where $classes(S_{T'}) = \{n_{11}, n_{12}\}$, $classes(S \setminus S_{T'}) = \{n_{21}, n_{22}\}$ and $n = |S| = n_{11} + n_{12} + n_{21} + n_{22}$:

	t_c	$t_{\bar{c}}$
$S_{T'}$	n_{11}	n_{12}
$S \setminus S_{T'}$	n_{21}	n_{22}

Fisher's exact test then evaluates the probability of obtaining the observed distribution (counts n_{ij}), or a more extreme one, assuming that the marginal counts (t_c, $t_{\bar{c}}$, $S_{T'}$, $S \setminus S_{T'}$) are fixed [8]. Therefore, first, the probability of observed quantities is calculated by

$$P(X = n_{11}) = \binom{n_{11} + n_{12}}{n_{11}} \binom{n_{21} + n_{22}}{n_{21}} \Big/ \binom{n}{n_{11} + n_{21}}. \tag{5}$$

Then the p-value is the sum of all probabilities for the observed or more extreme (that is, $X < n_{11}$) observations:

$$p = \sum_{i=0}^{n_{11}} P(X = i). \tag{6}$$

Example 2.6. Consider the bank customers in Table 1, the time point before the financial crisis, the attribute value set $T' = \{village\}$ and the classes $t_c = high$ vs. $t_{\bar{c}} = low$ for the class attribute *Spending*. There are two bank customers living in a *village*: $S_t = S_{village} = \{19, 20\}$, of which none is annotated by $Spending = high$. Hence, Fisher's exact p-value is $p \approx 0.237$.

Permutation test. In our experiments, to address the multiple testing problem, we perform a simple permutation test that returns adjusted p-values (see Appendix for the details).

Subgroup discovery. We can now describe the problem of subgroup discovery formally.

Definition 2.2. *The* subgroup discovery problem *is to output all sets $T' \subset T$ of attribute values for which $f_c(T') \leq \alpha$ for some given constant α.*

Equivalently, the subgroups defined by the sets of attribute values could be output, and in practice both, the sets of attribute values and subgroups are often shown as a result. An alternative formulation of the problem is to output the k best subgroups instead of using a fixed threshold.

Example 2.7. Consider again the bank customers in Table 1. When using Fisher's exact test, the adjusted p-value as class-related interestingness measure $f_c(\cdot)$, and $\alpha = 0.3$ (for the sake of simplicity, we consider a relatively high threshold in this toy example), a subgroup discovery method finds four interesting subgroups for the time point before the financial crisis: $village \mapsto \{19, 20\}$, $unemployed \mapsto \{13, 17, 19, 20\}$, $unemployed \wedge city \mapsto \{13, 17\}$ and $unemployed \wedge village \mapsto \{19, 20\}$ as well as two interesting subgroups $education \mapsto \{6, 9, 11, 16, 18\}$ and $education \wedge city \mapsto \{6, 9, 11, 16, 18\}$ for the time point after the financial crisis.

2.2. Other pattern mining approaches

Other pattern mining approaches mentioned below can be classified as unsupervised and supervised. Unsupervised methods (frequent item set mining and association rule mining) take a dataset without class labels as input, while the input to supervised methods (the other methods listed below) is a class labeled dataset. Note that the supervised methods can take multiple classes into account by comparing two classes where one is a union of several (sub)classes [9].

Frequent item set mining aims to find frequent combinations of attribute values (items) such as $Occupation = industry \wedge Spending = high$ [10]. Similar to the approach presented here, some methods intersect transactions to find closed frequent item sets [11–13].

Emerging patterns are item sets for which the supports increase significantly from one class to another [14].

Association rules describe associations, such as $X \mapsto Y$, where the antecedent X and consequent Y are item sets (e.g. sets of terms) [15]. In categorical data, the antecedent and consequent are (attribute, attribute value) pairs such as $Occupation = industry \mapsto Spending = high$ [16, 17].

Exception rule mining aims to find unexpected association rules that differ from a highly frequent association rule [18]. That is, it finds unexpected association rules $X \wedge Z \mapsto Y$, where $X \mapsto Y'$ and $Z \not\mapsto Y'$. Here, X and Z are item sets or (attribute, attribute value) pairs, and Y and Y' are different (class attribute, class) pairs. Consider, for example, X as $Occupation = industry$, Z as $Location = city$, Y as $Spending = high$ and Y' as $Spending = low$.

Contrast set mining is an extension of association rule mining and aims to understand the differences between contrasting groups of objects [16, 17, 19, 20]. Contrast set mining and emerging pattern mining are formally equivalent and can be effectively solved by subgroup discovery methods [17, 21]. In contrast set mining two contrast classes are defined, while in subgroup discovery only one class and its complement are used.

Examples of contrast set mining methods are Search and Testing for Understandable Consistent Contrasts [16], Contrasting Grouped Association Rules [20] and Rules for Contrast Sets [22] all of which derive rules of attribute-value pairs for which the support differs meaningfully across groups.

In a setting where several different class attributes exist, these methods can be applied in a pairwise manner. For example, one could contrast two different levels of spending for different time points or different locations separately. That is, these methods find rules such as $Occupation = industry \wedge Spending = high$ for which the support is significantly larger within the individuals that are described by $Location = city$ than $Location = village$.

We also aim to understand the differences between several contrasting groups. However, in contrast to contrast set mining and the other approaches described here, our aim is to find interesting subgroups of objects which are characteristic for their class, regardless of their class. Next we describe the problem formally.

3. PROBLEM DEFINITION

We now formulate the problem of contrasting subgroup discovery in more exact terms. We replace the direct dependency on the class distribution of the classical subgroup discovery by a contrasting, indirect one. In the classical, direct case, one is interested in sets of attribute values that are characteristic for a class. Our aim to understand phenomena in a setting where several different classes (for example, different time points) are given. That is, in the contrasting case, we want to find sets of attribute values that indicate objects that are characteristic for their class, but not necessarily the same one.

To formally define the task, we first introduce a notation P for the set of objects characteristic for their class:

$$\begin{aligned} P = \{s \in S \mid {} & \text{there exists } T' \subset T \text{ such that} \\ & f_{\mathrm{c}}(T') \leq \alpha \text{ and } s \in S_{T'}\}, \end{aligned} \tag{7}$$

where (as before) T denotes the set of attribute values (e.g. terms), S the set of objects, $S_{T'}$ the set of objects annotated by the attribute value set $T' \subset T$, $f_{\mathrm{c}}(\cdot)$ the class-related interestingness measure and α a given constant.

EXAMPLE 3.1. Consider again the bank customers in Table 1 and the subgroups found with a classical subgroup discovery method (see Example 2.7). Then the set of objects characteristic for their class is $P = \{13, 17, 19, 20\}$ for the time point before and $P = \{6, 9, 11, 16, 18\}$ for the time point after the financial crisis.

Now the user can define two contrast classes $P_{\mathrm{c}}, \overline{P_{\mathrm{c}}} \subset P$. The selection of these two contrast classes depends on the objective and is left to the user. They can, for example, take several classes (such as different time points) into account.

Let $c_1, \ldots, c_m$ be m class attributes and $P_1, \ldots, P_m$ be the sets of objects characteristic for each of the class attributes. Here, we define P_{c} and $\overline{P_{\mathrm{c}}}$ in two different, exemplary ways. First, P_{c} can be defined as the set of objects occurring in interesting subgroups of *all* class attributes:

$$P_{\mathrm{c}} = \bigcap_{i \in \{1,\ldots,m\}} P_i. \tag{8}$$

This is useful when one wants to find interesting subgroups that are common to all class attributes (for example, a specific time point in contrast to all other time points).

Secondly, P_{c} can be defined as the set of objects occurring *only* in interesting subgroups of the kth class attributes:

$$P_{\mathrm{c}} = P_k \setminus \bigcup_{\substack{i \in \{1,\ldots,m\},\\ i \neq k}} P_i. \tag{9}$$

This definition can be used to find interesting subgroups that are specific for one class attribute in contrast to all the other class attributes.

The contrast class $\overline{P_{\mathrm{c}}}$ can be defined as the complement of P_{c}, that is,

$$\overline{P_{\mathrm{c}}} = P \setminus P_{\mathrm{c}}, \tag{10}$$

when one is interested in subgroups specific for the objects in P_{c} compared with all other objects of P. Or, if a user is interested in contrasting two specific time points even in a case where more time points exist. Then P_{c} would be defined as one of those time points and $\overline{P_{\mathrm{c}}}$ as the other time point.

EXAMPLE 3.2. In the case of bank customers and the two classes before and after the financial crisis, we obtain the set of objects characteristic for each class attribute separately, that is, $P_1 = \{13, 17, 19, 20\}$ and $P_2 = \{6, 9, 11, 16, 18\}$. When specifying the contrast classes P_{c} and $\overline{P_{\mathrm{c}}}$ as $P_{\mathrm{c}} = P_1 \setminus P_2$ and $\overline{P_{\mathrm{c}}} = P_2$ (Equations (9) and (10)), we contrast the time point before the financial crisis against the time point after the financial crisis and obtain $P_{\mathrm{c}} = \{13, 17, 19, 20\}$ as well as $\overline{P_{\mathrm{c}}} = \{6, 9, 11, 16, 18\}$ as also shown in Table 2. (Note that we could alternatively contrast the time point after against the time point before the financial crisis by defining $P_{\mathrm{c}} = P_2 \setminus P_1$ and $\overline{P_{\mathrm{c}}} = P_1$.)

Let us define function $characteristic(\cdot)$ that gives the number of objects characteristic for their class in the contrasting classes P_{c} and $\overline{P_{\mathrm{c}}}$ for a given set $S_{T'}$:

$$\begin{aligned} characteristic : {} & \mathcal{P}(S) \to \mathbb{Z}_+ \times \mathbb{Z}_+, \\ & S_{T'} \mapsto (|S_{T'} \cap P_{\mathrm{c}}|, |S_{T'} \cap \overline{P_{\mathrm{c}}}|). \end{aligned} \tag{11}$$

Now, the contrasting interestingness measure, as well as the contrasting subgroup discovery problem, can be formulated as follows.

DEFINITION 3.1. *A* contrasting interestingness measure *is a function*

$$\begin{aligned} f_i : {} & \mathcal{P}(T) \to \mathbb{R}, \\ & T' \mapsto g'(characteristic(S_{T'}), \\ & \qquad characteristic(P \setminus S_{T'})), \end{aligned} \tag{12}$$

for some function g'.

That is, the contrasting interestingness measure analyzes whether a subgroup is interesting w.r.t. the two contrast classes, which both consist only of objects that are characteristic for their own class. This is in contrast to the classical class-related

TABLE 2. Contrast classes of bank customers.

ID	Occupation	Location	Contrast class
6	Education	Big city	$\overline{P_{\mathrm{c}}}$
9	Education	Small city	$\overline{P_{\mathrm{c}}}$
11	Education	Big city	$\overline{P_{\mathrm{c}}}$
13	Unemployed	Big city	P_{c}
16	Education	Small city	$\overline{P_{\mathrm{c}}}$
17	Unemployed	Small city	P_{c}
18	Education	Small city	$\overline{P_{\mathrm{c}}}$
19	Unemployed	Village	P_{c}
20	Unemployed	Village	P_{c}

interestingness measure, which analyzes whether a subgroup is interesting w.r.t. the object's classes.

Example 3.3. Consider again the bank customers and the two contrast classes $P_c = \{13, 17, 19, 20\}$ and $\overline{P_c} = \{6, 9, 11, 16, 18\}$ of Table 2. Then the attribute value *Occupation* = *education*, for instance, defines a set of five bank customers $\{6, 9, 11, 16, 18\}$, who are all in $\overline{P_c}$. Given the adjusted p-value as function g', we obtain $f_i(education) \approx 0.0079$.

Definition 3.2. *The* contrasting subgroup discovery problem *is to output all sets $T' \subset T$ of attribute values for which $f_i(T') \leq \alpha'$ for some given constant α'.*

In other words, while classical subgroup discovery is related to the question of how to find sets of objects that are characteristic for a specific class, the problem of contrasting subgroup discovery is related to asking if sets of objects characteristic for (any) classes can be found.

The relationship between the classical and contrasting cases immediately implies that, for any subgroup found for the contrasting subgroup discovery problem, its objects are characteristic for their class. On the other hand, a set of attribute values may be a valid answer to the contrasting problem even if it is not for the classical problem.

That is exactly where the main conceptual contribution of this paper is. Contrast subgroup discovery allows finding subgroups of objects that could not be found with classical subgroup discovery.

4. METHOD

Given a set of objects described by attribute values (e.g. terms) and different classes of objects, our goal is to find interesting subgroups of objects characteristic for their class. Thereby we allow taking different class attributes into account.

To find such subgroups, we propose an approach that consists of three steps: First, interesting subgroups are found by a classical subgroup discovery method. Secondly, contrast classes on those subgroups are defined by set-theoretic functions. Thirdly, contrasting subgroup discovery finds interesting subgroups in the contrast classes. Next, we will describe each step in detail.

Classical subgroup discovery (*Step* 1). Given some objects that are annotated by attribute values, and assigned a class, a subgroup discovery method is applied. Thereby, we consider only one class attribute [e.g. Spending before the financial crisis with different classes (e.g. $Spending = high$ vs. $Spending = low$)], and apply a subgroup discovery method separately for each class attribute (e.g. separately for each time point). The subgroups are then analyzed by a statistical test, like Fisher's exact test followed by a permutation test. (See Example 2.7 for exemplary results of classical subgroup discovery.)

Construction of contrast classes (*Step* 2). Let $P_1, \ldots, P_m$ denote the objects characteristic for their class of m class attributes (e.g. for m different time points). Then the two contrast classes P_c and $\overline{P_c}$ are defined by two set-theoretic functions, for example, by Equations (9) and (10). (As stated before, the selection of a particular set-theoretic function depends on the objective and is left to the user.)

Contrasting subgroup discovery (*Step* 3). In this step, we apply a second subgroup discovery instance in order to analyze subgroups with respect to the constructed contrast classes. Given the objects in the two contrast classes P_c and $\overline{P_c}$, we find interesting subgroups of these objects by a second subgroup discovery instance. Again, the p-values are calculated, using a permutation test.

Assuming that both subgroup discovery instances (Steps 1 and 3) find all subgroups for which the classical interesting measures hold (Equation (4)), the proposed method does find all subgroups that satisfy the indirect interestingness measure (Equation (12)).

Example 4.1. In the case of bank customers, we saw already in Example 3.3 that *education* is obtained with contrasting subgroup discovery when the two classes $P_c = \{13, 17, 19, 20\}$ and $\overline{P_c} = \{6, 9, 11, 16, 18\}$ are contrasted. In this contrasting subgroup, discovery the following subgroups are found to be interesting:

$$\begin{aligned}
education &\mapsto \{6, 9, 11, 16, 18\},\\
education \wedge city &\mapsto \{6, 9, 11, 16, 18\},\\
education \wedge big\ city &\mapsto \{6, 11\},\\
education \wedge small\ city &\mapsto \{9, 16, 18\},\\
public &\mapsto \{6, 9, 11, 16, 18\},\\
public \wedge city &\mapsto \{6, 9, 11, 16, 18\},\\
public \wedge big\ city &\mapsto \{6, 11\}
\end{aligned}$$

and

$$public \wedge small\ city \mapsto \{9, 16, 18\}.$$

In contrast, with a classical subgroup discovery method we obtain

$$\begin{aligned}
village &\mapsto \{19, 20\},\\
unemployed &\mapsto \{13, 17, 19, 20\},\\
unemployed \wedge city &\mapsto \{13, 17\}
\end{aligned}$$

and

$$unemployed \wedge village \mapsto \{19, 20\},$$

for the time point before the financial crisis and

$$education \mapsto \{6, 9, 11, 16, 18\}$$

and

$$education \wedge city \mapsto \{6, 9, 11, 16, 18\},$$

for the time point after the financial crisis.

Hence, some of the subgroups found by the contrasting subgroup discovery were already found by the classical subgroup discovery (for example, $education \wedge city$). Other subgroups found by the contrasting subgroup discovery are more specific than the one found by the classical subgroup discovery (for example, $education \wedge big\ city$). Again other subgroups found by the contrasting subgroup discovery were not found at all by the classical subgroup discovery (for example, $public \wedge big\ city$) as its members are not characteristic for either class (that is, some of them are assigned the class $Spending = high$ and some $Spending = low$).

Both, more specific and new subgroups might reveal new research hypotheses for the user. For example, $public \wedge big\ city$ defines in classical subgroup discovery a subgroup that is not interesting since its objects are characteristic for either *high* or *low* spending. In contrasting subgroup discovery, it defines a subgroup that is characteristic when the two contrasting classes are analyzed. That is, this subgroup's objects occur only in subgroups that are characteristic for the time point after the financial crisis, but not in one that is characteristic for the time point before the financial crisis. Hence, there has been some changes in those subgroups between the two points. This directs the user where to look for the causes of the differences between the time points. Other methods (and possibly data) are needed to find those causes.

5. AN APPLICATION IN BIOLOGY

Application areas of subgroup discovery include sociology [1, 2], marketing [23], vegetation data [24] and transcriptomics [25]. In bioinformatics, high-throughput techniques and simple statistical tests are used to produce rankings of thousands of genes. Life-scientists have to choose a few genes for further (often expensive and time consuming) experiments. In this context, subgroup discovery is known as gene set enrichment (see, e.g. [26, 27]).

A lifescientist might be interested in studying an organism in virus-infected and non-infected conditions at different time points or in different phenotypes of that organism. Here, our aim is to find enriched gene sets characteristic for their class, regardless of their class (for example, characteristic for either differently expressed or not). Further, we allow the user to specify subsets of objects she wants to contrast. The lifescientist could then specify that she is interested in objects at a specific time point in contrast to all other, or for a specific phenotype in contrast to all other phenotypes. With our proposed approach of contrasting subgroup discovery, we can then contrast several time points or phenotypes at the same time.

Using subgroup discovery terminology, we consider genes as objects, and their annotations by terms (e.g. by their molecular functions or biological processes) as attribute values. Table 3 aligns the terms used in the data mining and bioinformatics communities to provide a better understanding of the terminologies.

TABLE 3. Synonyms from different communities.

Subgroup discovery	Bioinformatics
Object or instance	Gene
Attribute value or feature value, e.g. a term in a hierarchy	Annotation or biological concept, e.g. a GO term
Class attribute	Gene expression under a specific experimental condition such as a specific time point or phenotype
Class or class attribute value, e.g. positive/negative	Differential/non-differential gene expression
Subgroup of objects	Gene set
Interesting subgroup	Enriched gene set

Next, we describe measures used for transforming the expression values of several samples (e.g. virus infected vs. non-infected plants) into a class attribute, called differential expression, how the constructed gene sets are analyzed for statistical significance and how enriched gene sets can be found. Finally, we discuss how our proposed method finds contrasting gene sets.

5.1. Measures of differential expression

After preprocessing the gene expression data (including microarray image analysis and normalization), the genes can be ranked according to their gene expression. The dataset of our experiments consists of four samples for both experimental conditions. That is, for each gene we have gene expression levels for four replicates of virus-infected and for four replicates of non-infected plants. Different methods can be used to transform several samples into one class attribute. Here, we will discuss two widely used ones.

Fold change (*FC*) is a metric for comparing the expression level of a gene g between two distinct experimental conditions, for example, virus-infected and non-infected [25]. FC is defined as the log ratio of the average gene-expression levels with respect to the two conditions [28]. Note that FC values do not indicate the level of confidence in the designation of genes as differently expressed or not.

The *t-test* is used to determine the statistical significance of the gene expression between two distinct experimental conditions [25] though, the power of the test is relatively low for small sample sizes [28]. A Bayesian t-test is advantageous if only a few (that is, two or three) samples are used, but no advantage is gained if more replicates are used [29]. In our experiments, we use four replicates and therefore will use the simple t-test.

5.2. Analysis of gene set enrichment

Given a list $L = \{g_1, \dots, g_n\}$ of n genes in which all genes of S are ranked by their expression levels $e_1, \dots, e_n$, we can analyze the enrichment of a gene set $S_{T'}$ compared with the other genes $S \setminus S_{T'}$ with statistical tests like Fisher's exact test [8]. Alternatively, gene set enrichment analysis (GSEA) [30] or parametric analysis of gene set enrichment (PAGE) [27] can be used. Both methods use the ranking of differential expressions, instead of a partition of the genes into two classes.

Fisher's exact test. When analyzing the gene set $S_{T'}$ compared with the other genes $S \setminus S_{T'}$ with Fisher's exact test, we need to divide the genes into two classes t_c and $t_{\bar{c}}$. Therefore, a cut-off is set in the gene ranking: genes in the upper part are defined as differentially expressed and the genes in the lower part are defined as not differentially expressed genes. Then the p-values are calculated and a permutation test is performed.

GSEA evaluates whether objects of $S_{T'}$ are randomly distributed throughout the list L or primarily found at the top or bottom of the list [26, 30]. An enrichment score (ES) is calculated, which is the maximum deviation from zero of the fraction of genes in the set $S_{T'}$ weighted by their correlation and the fraction of genes not in the set:

$$\mathrm{ES}(S_{T'}) = \max_{i\in\{1,\dots,n\}} \left| \sum_{\substack{g_j\in S_{T'}\\ j\le i}} \frac{|e_j|^p}{n_w} - \sum_{\substack{g_j\notin S_{T'}\\ j\le i}} \frac{1}{n-n_w} \right|, \qquad (13)$$

where $n_w = \sum_{g_j\in S_{T'}} |e_j|^p$. If the ES is small, then $S_{T'}$ is randomly distributed across L. If it is high, then the genes of $S_{T'}$ are concentrated in the beginning or the end of the list L. The exponent p controls the weight of each step. We see that $\mathrm{ES}(S_{T'})$ reduces to the standard Kolmogorov–Smirnov statistic if $p = 0$:

$$\mathrm{ES}(S) = \max_{i\in\{1,\dots,n\}} \left| \sum_{\substack{g_j\in S_{T'}\\ j\le i}} \frac{1}{|S_{T'}|} - \sum_{\substack{g_j\notin S_{T'}\\ j\le i}} \frac{1}{|S|-|S_{T'}|} \right|. \qquad (14)$$

The significance of $\mathrm{ES}(S_{T'})$ is then estimated by a permutation test.

PAGE is a GSEA method based on a parametric statistical analysis model [27]. For each gene set, $S_{T'}$ a Z-score is calculated, which is the fraction of the mean deviation to the standard deviation of the ranking score values:

$$Z(S_{T'}) = (\mu_{S_{T'}} - \mu)\frac{1}{\sigma}\sqrt{|S_{T'}|}, \qquad (15)$$

where σ is the standard deviation, and μ and $\mu_{S_{T'}}$ are the means of the score values for all genes and for the genes in set $S_{T'}$, respectively. The Z-score is high if the deviation of the score values is small or if the means largely differ between the gene set and all genes. As gene sets may vary in size, the fraction is scaled by the square root of the set size. However, because of this scaling the Z-score is also high if $S_{T'}$ is very large. Assuming a normal distribution, a p-value for each gene set is calculated. Finally, the p-values are corrected by a permutation test.

Kim and Volsky [27] studied different datasets for which PAGE generally detected a larger number of significant gene sets than GSEA. On the other hand, GSEA makes no assumptions about the variability and can be used if the distribution is not normal or is unknown.

Trajkovski *et al.* [7] used the sum of GSEA's and PAGE's p-values, weighted by percentages (e.g. one-third of GSEA's and two-thirds of PAGE's or half of both). Hence, gene sets with small p-values for GSEA and PAGE are output as enriched gene sets.

5.3. Finding enriched gene sets with searching for enriched gene sets

In our experiments, we use the searching for enriched gene sets (SEGS) method [7] to find interesting subgroups of objects (that is, enriched gene sets). There, a subgroup of objects is considered interesting, when the subgroup is large enough, and its p-value obtained by a statistical test is smaller than the given significance level α.

SEGS uses hierarchies of attribute values (here, terms) to construct subgroups by individual terms as well as by logical conjunctions of terms. Ontologies are extensively used in gene set enrichment [27, 30]. Commonly used ontologies include gene ontology[1] (GO) [31], KO[2] [Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology] [32] and GoMapMan[3], an extension of the MapMan [33] ontology, for plants.

SEGS combines terms from the same level as well as from different levels into term conjunctions as follows. Several ontologies can be modeled by a single ontology [34]. To construct all possible subgroups, one merged ontology is used, where the root has n children, one for each individual ontology. We start with the root term and recursively replace each term by each of its children.

We are not interested in constructing all possible subgroups, but only those representing at least a minimal number min of objects. This parameter min is specified by the user. We conjunctively extend a rule condition only if the subgroup defined by it contains more than a minimum number of objects. If a condition defines the same group of objects as a more general condition, the more general condition is deleted. Further, in each recursive step we add other terms to the rule condition to obtain intersections of two or more subgroups.

5.4. Finding contrasting gene sets

To find contrasting gene sets, that is, to find enriched gene sets (interesting subgroups) that are characteristic for their class, we can apply our proposed method described in Section 4.

[1] http://www.geneontology.org/.
[2] http://www.genome.jp/kegg/ko.html.
[3] http://www.gomapman.org/.

Note that there are a couple of issues to take into account in the case of gene set enrichment. In Step 1, the classical subgroup discovery, the subgroups can be analyzed by a statistical test, like Fisher's exact test followed by a permutation test or alternatively by GSEA and PAGE in the case of a gene set enrichment application. In Step 2, the user can then choose to contrast different time points or different phenotypes. In Step 3, the contrasting subgroup discovery, we need to analyze the constructed gene sets by a statistical test, like the Fisher's exact test. There, GSEA and PAGE cannot be used for analyzing the constructed gene sets since we analyze the subgroups with respect to two classes P_c and $\overline{P_c}$ (and not with respect to the differential expression which would provide a ranking for GSEA and PAGE).

6. EXPERIMENTS AND RESULTS

For our experiments, we used a *S. tuberosum* (potato) time labeled gene expression dataset for virus-infected and non-infected plants. *S. tuberosum* is severely damaged by the *potato virus Y*. When infected, the plant shows severe symptoms within 1 week and dies after several weeks. Biologists aim to understand the plants disease response by utilizing gene set enrichment.

The dataset consists of three time points: 1, 3 and 6 days after virus infection when the viral-infected leaves as well as leaves from non-infected plants were collected. The aim is to find enriched gene sets that are common to virus-infected plants compared with non-infected plants and at the same time specific for one or all time points. Hence, we transform the expression values of our four samples (four virus-infected and four non-infected plants) into a class attribute, called differential expression, for each time point separately (see Section 5 for details). Afterward we have three class attributes, one for each time point, and can apply our proposed contrasting subgroup discovery method to contrast the different time points.

Recently, *S. tuberosum*'s genome has been completely sequenced [35], but only a few GO or KEGG annotations of *S. tuberosum* genes exist. However, plenty of GO and KEGG annotations exist for the well-studied model plant *Arabidopsis thaliana*. Therefore, we perform two approaches: First, we use homologs between *S. tuberosum* and *A. thaliana* and ontologies for *A. thaliana*. Second, we build *S. tuberosum* ontologies using homolog sequences in the National Center for Biotechnology Information (NCBI) and their GO annotations. For both approaches, we carried out gene set enrichment experiments in an Orange4WS[4] workflow [36].

Our interest is in assisting biologists to generate new research hypotheses. Therefore, we evaluate our results by counting the quantities of gene sets which are unexpected as well as those that are useful to a plant biologist (as in [37]). In this context, *unexpected* means that the knowledge was contained in GO, KEGG or GoMapMan, but it was not shown previously to be related to *S. tuberosum*'s response to viral infection. A gene set is *useful* if it is of interest for the plant biologist, that is, the gene set description tells him something about the virus response, and/or he might want to have a closer look at the genes of that gene set. We compare the results obtained by our proposed method (Steps 1–3) to those results obtained with a classical subgroup discovery method (Step 1).

6.1. *A. thaliana* homologs approach

Experimental setting. We use homologs between *S. tuberosum* and *A. thaliana* to make GSEA for *S. tuberosum* possible. There are more than 26 000 homologs for more than 42 000 *S. tuberosum* genes. GSEA is performed based on expression values in the dataset, the gene IDs of the *A. thaliana* homologs, and GO and KEGG annotations for *A. thaliana*.

We restricted gene sets to contain a minimum of three genes ($min = 3$) as only these are biologically relevant, the gene set description to contain a maximum of four terms, and the p-value to be 0.05 or smaller. For analyzing the constructed gene sets obtained by classical subgroup discovery (Step 1), we used Fisher's exact test, GSEA, PAGE and the combined GSEA and PAGE with equal percentages. Fisher's exact test was used to analyze gene set enrichment obtained by contrasting subgroup discovery (Step 3).

We considered two types of contrast classes for gene set enrichment (Step 2). First, the intersection: genes that are common to all classes compared with the genes occurring in some gene sets, but not in all (obtained by Equation (8)). Second, the set differences: genes that are specific for one class compared with the genes of the gene sets of the other classes (obtained by Equation (9)). The choice was made by the plant biologists, who are interested in understanding which biological processes, pathways, etc. are active at all time points, and which are active only at a specific time point.

Results. The quantities of enriched gene sets found with the *A. thaliana* homolog approach are shown in Table 4. The first subgroup discovery instance (Step 1), that is, the classical subgroup discovery method, found only a few, if any, gene sets. All gene sets that were found for the classical subgroup discovery method (Step 1) are described by either

```
protein.synthesis.ribosomal protein.
prokaryotic (GoMapMan:29.2.1.1)
```

more general terms of this gene set description (that is, for example, by `GoMapMan:29.2.1`), or by

```
Plant-pathogen interaction (KEGG:04626)
```

As we construct the set differences and intersection from the gene sets found in Step 1, it is no surprise that also the second subgroup discovery instance (Step 3), the contrasting subgroup discovery method, found only a few, if any, gene sets at all. Some of the gene sets that were found by the contrasting subgroup

[4] http://orange4ws.ijs.si/.

TABLE 4. Quantities of enriched gene sets found with the classical subgroup discovery (Step 1) and with the contrasting subgroup discovery method (Step 3) for the *A. thaliana* homologs approach with Fisher (F), GSEA (G), PAGE (P) and GSEA and PAGE combined (C).

	F	G	P	C
Classical SD (Step 1)				
Day 1	1	0	2	0
Day 3	0	0	1	0
Day 6	1	0	0	0
Contrasting SD (Step 3)				
Day 1 set difference	6	0	0	0
Day 3 set difference	0	0	0	0
Day 6 set difference	3	0	0	0
Intersection	0	0	0	0

discovery method (Step 3) are more specific than those found by the classical subgroup discovery method (Step 1). For instance,

```
protein.synthesis.ribosomal protein.
prokaryotic.chloroplast.50S subunit
(GoMapMan:29.2.1.1.1.2)
```

is more specific than `GoMapMan:29.2.1.1`, which is a term located higher in the term hierarchy. Another example is

```
calmodulin-dependent protein kinase activity
(GO:0004683)
∧ signalling.calcium (GoMapMan:30.3)
∧ Plant-pathogen interaction (KEGG:04626)
```

where `KEGG:04626` became combined with terms from other hierarchies. This combination was not statistically significant for the classical subgroup discovery method (Step 1), but is for the contrasting subgroup discovery method (Step 3), when comparing the contrast sets constructed in Step 2.

No gene sets at all were unexpected using the *A. thaliana* homologs approach. A gene set is *useful* if it is of interest for the plant biologist, that is, the gene set description tells him something about the virus response, and/or he might want to have a closer look at the genes of that gene set. The quantities of unexpected enriched gene sets found using the *A. thaliana* homologs approach are shown in Table 5. The only gene set that is useful for the classical subgroup discovery method (Step 1) is

```
Plant-pathogen interaction (KEGG:04626)
```

which covers 51 genes with a p-value $\leq 10^{-6}$. This gene set description is expected as it describes the plant's defense pathway to disease infections.

Two enriched gene sets were found to be useful for the contrasting subgroup discovery method (Step 3) on the first day:

```
protein.synthesis.ribosomal protein.
prokaryotic.chloroplast
(GoMapMan:29.2.1.1.1)
```

which covers 28 genes with a p-value $\leq 10^{-6}$ and its more specific variant

```
protein.synthesis.ribosomal protein.
prokaryotic.chloroplast.50S subunit
(GoMapMan:29.2.1.1.1.2)
```

which covers 22 genes with a p-value $\leq 10^{-6}$. The more general as well as the more specific gene set description was output as they define different gene sets. More precisely, the gene set of the more specific description is a subset of the gene set of the more general description.

TABLE 5. Quantities of *useful* enriched gene sets found with the *A. thaliana* homologs approach.

	F	G	P	C
Classical SD (Step 1)				
Day 1	0	0	0	0
Day 3	0	0	0	0
Day 6	1	0	0	0
Contrasting SD (Step 3)				
Day 1 set difference	2	0	0	0
Day 3 set difference	0	0	0	0
Day 6 set difference	2	0	0	0
Intersection	0	0	0	0

For the contrasting subgroup discovery (Step 3), only enriched gene sets are counted that are useful as well as new or more specific in comparison to the classical subgroup discovery (Step 1).

The two enriched and useful gene sets found for the contrasting subgroup discovery method (Step 3) on Day 6 are

```
Plant-pathogen interaction (KEGG:04626)
∧ signalling.calcium (GoMapMan:30.3)
```

which covers 26 genes with a p-value $\leq 10^{-6}$, and

```
Plant-pathogen interaction (KEGG:04626)
∧ signalling.calcium (GoMapMan:30.3)
∧ Calmodulin-dependent protein kinase
activity (GO:0004683)
```

which is more specific than the previous one, and covers only 14 genes with a p-value of 0.0001. All these gene sets are described by more specific concepts than those found with the classical subgroup discovery method (Step 1) and hence give the plant biologists more detailed information.

For the intersection in Step 3, we obtained no enriched gene sets at all. This reflects the characteristics of a defense response: The gene expression of the first days (when activating the defense response) differs from the gene expression on Day 6 (when the defense response is active) and therefore the intersection reveals no enriched gene sets that are active at all time points.

TABLE 6. Quantities of enriched gene sets found with the classical (Step 1) and with the contrasting subgroup discovery method (Step 3) for the *S. tuberosum* GO approach with Fisher (F), GSEA (G), PAGE (P), and GSEA and PAGE combined (C).

	F	G	P	C
Classical SD (Step 1)				
Day 1	7	2	5	2
Day 3	2	0	5	0
Day 6	3	1	12	1
Contrasting SD (Step 3)				
Day 1 set difference	16	3	15	3
Day 3 set difference	3	0	15	0
Day 6 set difference	29	2	29	2
Intersection	0	0	0	0

6.2. *S. tuberosum* GO approach

Experimental setting. We built *S. tuberosum* ontologies independently using Blast2GO[5] to obtain homolog sequences in NCBI and their GO annotations. Enrichment analysis is then performed using *S. tuberosum*'s gene IDs and expression values, and GO and KEGG annotations obtained with Blast2GO.

Again, we restricted gene sets to contain a minimum of three genes ($min = 3$), the gene set description to contain a maximum of four terms and the p-value to be 0.05 or smaller. For analyzing the constructed gene sets, we used the Fisher's exact test, GSEA, PAGE and the combined GSEA and PAGE with equal percentages, in Step 1, and Fisher's exact test in Step 3. We considered the same two types of contrast classes for gene set enrichment (Step 2) as in the *A. thaliana* approach: the intersection (Equation (8)) and the set differences (obtained by Equation (9)).

Results. The quantities of enriched gene sets found with the *S. tuberosum* GO approach are shown in Table 6. In comparison with the *A. thaliana* approach, we found more enriched gene sets. This is probably due to the following reason: Many potato genes have no homologs in *A. thaliana* or the homologs are not known yet, but with the *S. tuberosum* GO approach we obtain extensive GO annotation of the genes.

However, when using GSEA (either alone or in combination with PAGE) to analyze the constructed gene sets of the first subgroup discovery instance (Step 1), that is, the classical subgroup discovery method, only a few more enriched gene sets are found. When Fisher's exact test or PAGE are used instead, more enriched gene sets are found. This suggests that especially in the *S. tuberosum* gene ontology approach one of these methods should be preferred.

When PAGE is used, several enriched gene sets are found on Day 6 by the classical subgroup discovery method (Step 1). Even more enriched gene sets are found by the contrasting subgroup discovery method (Step 3) on Day 6 when PAGE or Fisher's exact test are used in Step 1. (As stated before, in Step 3 always Fisher's exact test is used to analyze the constructed gene sets.) The fact that more enriched gene sets are found on Day 6 reflects that *S. tuberosum* activates the defense response in the first days, and the full effect can be witnessed only on Day 6.

Several gene sets that are known to relate to *S. tuberosum*'s response to virus infection were found, including molecular functions, biological processes and pathways with a central role in it, such as

```
auxin mediated signalling pathway
(GO:0009734)
```

which covers 42 genes with a p-value $\leq 10^{-6}$,

```
fatty acid catabolic process (GO:0009062)
∧ lipid metabolism.lipid degradation.
beta-oxidation (GoMapMan:11.9.4)
```

which covers 17 genes with a p-value of 0.0001, and

```
protein.postranslational modification
(GoMapMan:29.4)
∧ protein serine/threonine phosphatase
complex (GO:0008287)
```

which covers 16 genes with a p-value of 0.0001.

As before, we counted the quantities of enriched gene sets that are unexpected to a plant biologist when using the *S. tuberosum* GO approach (see Table 7). In contrast to the *A. thaliana* approach, we found some enriched genes set that are unexpected. For the classical subgroup discovery method (Step 1), we found unexpected gene sets only on the first day, which all relate to the Golgi complex, such as

```
protein.targeting.secretory pathway.golgi
(GoMapMan:29.3.4.2)
```

which covers 19 genes with a p-value $\leq 10^{-6}$.

TABLE 7. Quantities of *unexpected* enriched gene sets found with the *S. tuberosum* GO approach.

	F	G	P	C
Classical SD (Step 1)				
Day 1	1	2	1	2
Day 3	0	0	0	0
Day 6	0	0	0	0
Contrasting SD (Step 3)				
Day 1 set difference	0	1	4	1
Day 3 set difference	0	0	0	0
Day 6 set difference	2	1	0	0
Intersection	0	0	0	0

For the contrasting subgroup discovery (Step 3), only enriched gene sets are counted that are unexpected as well as new or more specific in comparison to the classical subgroup discovery (Step 1).

[5] http://www.blast2go.org/.

For the contrasting subgroup discovery method (Step 3), we found unexpected gene sets for the first and sixth day. Some of those relate also to the Golgi complex, but were not found with the classical subgroup discovery method (Step 1), such as

```
ER to Golgi vesicle-mediated transport
(GO:0006888)
∧ vesicle coat (GO:0030120)
```

which covers 14 genes with a p-value of 0.0001. Other examples of unexpected gene sets are novel when compared with the enriched gene sets found by the classical subgroup discovery method (Step 1). Hence, they might reveal new research hypotheses for the plant biologists. Examples of such gene sets are

```
RNA.regulation of transcription.Chromatin
Remodeling Factors (GoMapMan:27.3.44)
```

which covers 15 genes with a p-value $\leq 10^{-6}$,

```
unidimensional cell growth (GO:0009826)
```

which covers 7 genes with a p-value of 0.0001 or

```
root development (GO:0048364)
∧ hormone metabolism.auxin (GoMapMan:17.2)
```

which covers 5 genes with a p-value of 0.001.

As before, we also counted the quantities of gene sets that are useful to a plant biologist when using the *S. tuberosum* GO approach (see Table 8).

An enriched gene set that was found by the classical subgroup discovery method (Step 1) and is considered useful is

```
protein.targeting.secretory pathway.golgi
(GoMapMan:29.3.4.2)
```

which covers 19 genes with a p-value $\leq$ 10^{-6}. Another example is

```
RNA.regulation of transcription.
WRKY domain transcription factor family
(GoMapMan:27.3.32)
```

which covers 30 genes with a p-value of 0.0001.

TABLE 8. Quantities of *useful* enriched gene sets found with the *S. tuberosum* GO approach.

	F	G	P	C
Classical SD (Step 1)				
Day 1	3	2	3	2
Day 3	0	0	3	0
Day 6	3	1	12	1
Contrasting SD (Step 3)				
Day 1 set difference	6	1	6	1
Day 3 set difference	1	0	6	0
Day 6 set difference	22	1	18	1
Intersection	0	0	0	0

For the contrasting subgroup discovery (Step 3), only enriched gene sets are counted that are useful as well as new or more specific in comparison with the classical subgroup discovery (Step 1).

Useful gene sets found by the contrasting subgroup discovery method (Step 3), which are novel or more specific when compared with the classical subgroup discovery method (Step 1) include

```
protein.degradation.ubiquitin.E3.SCF.FBOX
(GoMapMan:29.5.11.4.3.2)
```

which covers 40 genes with a p-value $\leq 10^{-6}$,

```
enoyl-CoA hydratase activity (GO:0004300)
```

which covers 7 genes with a p-value $\leq 10^{-6}$,

```
post-embryonic development (GO:0009791)
∧ reproductive structure development
(GO:0048608)
∧ RNA (GoMapMan:27)
```

which covers 21 genes with a p-value $\leq 10^{-6}$, and

```
ER to Golgi vesicle-mediated transport
(GO:0006888)
∧ vesicle coat (GO:0030120)
```

which covers 14 genes with a p-value of 0.0001. From these four gene sets, the last one is more specific, while the first three gene sets are novel when compared with the classical subgroup discovery method (Step 1).

Note that a gene set can be either expected and not useful, unexpected but not useful, expected, but useful, or both, unexpected as well as useful. Gene sets that are expected as well as not useful might be simply described by too general terms, such as

```
protein.postranslational modification
(GoMapMan:29.4)
```

which covers 217 genes with a p-value $\leq 10^{-6}$. A gene set can be expected and not useful also because it is not informative for some other reason, such as

```
coated vesicle membrane (GO:0030662)
```

which covers 27 genes with a p-value of 0.0001, but is not informative to plant biologists as it describes a cellular component only. An example of an enriched gene set that is unexpected, but not useful is

```
organ development (GO:0048513)
∧ RNA (GoMapMan:27)
```

which covers 21 genes with a p-value $\leq 10^{-6}$. It is not useful because the biological term it is too general.

Gene sets that are unexpected, useful or both may contain genes that are interesting for further (tough, time-consuming) wet-lab experiments. From the gene sets mentioned before, an example of an enriched gene set that is expected but useful is

```
RNA.regulation of transcription.
WRKY domain transcription factor family
(GoMapMan:27.3.32)
```

which is expected as it is known that these proteins have an important role in defense against virus, but still useful as it tells the plant biologist that these proteins are differentially expressed on Day 6. Examples of enriched gene sets that are unexpected and useful are

```
post-embryonic development (GO:0009791)
∧ reproductive structure development
(GO:0048608)
∧ RNA (GoMapMan:27)
```

and

```
ER to Golgi vesicle-mediated transport
(GO:0006888)
∧ vesicle coat (GO:0030120)
```

These rules combine two or more ontology terms that have not been associated with the viral infection response of plants (to the knowledge of the plant biologists). Therefore, the genes covered by these gene set descriptions are potentially interesting to the plant biologists and might help them to generate new hypotheses.

As in the *A. thaliana* approach, we did not obtain any enriched gene sets for the intersection in Step 3. Again, this reflects the characteristics of a defense response: The gene expression of the first days differs from the gene expression on Day 6.

7. CONCLUSIONS

We defined the problem of contrasting subgroup discovery; that is, the aim is to find subgroups of objects characteristic for their class, even if the classes are not identical. Further, we allow the user to specify contrast classes in which they are interested, for example, to contrast several time points. We proposed to find such subgroups by combining well-known algorithms. We showed that our approach finds subgroups of objects that are characteristic for their class, even if the classes are not identical. Our results on a time series dataset for virus-infected *S. tuberosum* (potato) plants indicate that such subgroups can be unexpected and useful for biologists. Studying the genes of such subgroups may reveal new research hypotheses for biologists.

Further experimental evaluation is planned, including an extensive evaluation of the quality of gene set descriptions which possibly relate to *S. tuberosum*'s virus response, but are unexpected for a plant biologist. Further, we will address the redundancy of gene set descriptions, and we will investigate how redundancy can be avoided, or at least decreased, for example, by rule clustering or filtering. In addition, we will evaluate the results at the gene level, including a selection of genes for wet-lab experiments, which will affect the understanding of the biological mechanisms of virus response, particularly that of *S. tuberosum*. Finally, we will perform further experiments on other, non-biological datasets and use simple as well as more complex set-theoretic functions.

ACKNOWLEDGEMENTS

We would like to thank Kamil Witek, Ana Rotter and Špela Baebler for the test data and the help with interpreting the results.

FUNDING

This work has been supported by the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open, contract no. BISON-211898, by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland and by the Slovenian Research Agency grants P2-0103, J4-2228 and P4-0165.

REFERENCES

[1] Klösgen, W. (1996) Explora: A Multipattern and Multistrategy Discovery Assistant. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, USA.

[2] Wrobel, S. (1997) An Algorithm for Multi-relational Discovery of Subgroups. In Komorowski, J. and Zytkow, J. (eds), *Principles of Data Mining and Knowledge Discovery*. Springer, Berlin.

[3] Bruner, J., Goodnow, J. and Austin, G. (1956) *A Study of Thinking*. Wiley, Hoboken, NJ, USA.

[4] Michalski, R. (1983) A theory and methodology of inductive learning. *Artif. Intell.*, **20**, 111–161.

[5] Gruber, T. (1995) Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, **43**, 907–928.

[6] Weber, I. (2000) Levelwise search and pruning strategies for first-order hypothesis spaces. *J. Intell. Inf. Syst.*, **14**, 217–239.

[7] Trajkovski, I., Lavrač, N. and Tolar, J. (2008) SEGS: search for enriched gene sets in microarray data. *J. Biomed. Inform.*, **41**, 588–601.

[8] van Belle, G., Fisher, L., Heagerty, P. and Lumley, T. (1993) *Biostatistics: A Methodology for the Health Sciences*. Wiley, Hoboken, NJ, USA.

[9] Li, J., Liu, G. and Wong, L. (2007) Mining statistically important equivalence classes and delta-discriminative emerging patterns. *Proc. KDD '07*, San Jose, CA, USA, August 12–15, pp. 430–439. ACM Press, New York City, NY, USA.

[10] Agrawal, R., Imieliński, T. and Swami, A. (1993) Mining Association Rules Between Sets of Items in Large Databases. *Proc. SIGMOD '93*, Washington, DC, USA, May 26–28, pp. 207–216. ACM Press, New York City, NY, USA.

[11] Mielikäinen, T. (2003) Intersecting Data to Closed Sets with Constraints. *Proc. FIMI '03*, Melbourne, FL, USA, Novermber 19, CEUR-WS.org, http://ceur-ws.org/Vol-90/mielikainen.pdf.

[12] Pan, F., Cong, G., Tung, A., Yang, J. and Zaki, M. (2003) Carpenter: Finding Closed Patterns in Long Biological Datasets.

Proc. KDD '03, Washington, DC, USA, August 24–27, pp. 637–642. ACM Press, New York City, NY, USA.

[13] Borgelt, C., Yang, X., Nogales-Cadenas, R., Carmona-Saez, P. and Pascual-Montano, A. (2011) Finding Closed Frequent Item Sets by Intersecting Transactions. *Proc. EDBT/ICDT '11*, Uppsala, Sweden, March 21–25, pp. 367–376. ACM Press, New York City, NY, USA.

[14] Dong, G. and Li, J. (1999) Efficient Mining of Emerging Patterns: Discovering Trends and Differences. *Proc. KDD '99*, pp. 43–52. ACM Press, New York City, NY, USA.

[15] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. (1996) Fast Discovery of Association Rules. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, USA.

[16] Bay, S. and Pazzani, M. (2001) Detecting group differences: mining contrast sets. *Data Min. Knowl. Discov.*, **5**, 213–246.

[17] Böttcher, M. (2011) Contrast and change mining. *Data Min. Knowl. Discov.*, **1**, 215–230.

[18] Suzuki, E. (1997) Autonomous Discovery of Reliable Exception Rules. *Proc. KDD '97*, Newport Beach, CA, USA, August 14–17, pp. 259–262. AAAI Press, Menlo Park, CA, USA.

[19] Webb, G., Butler, S. and Newlands, D. (2003) On Detecting Differences between Groups. *Proc. KDD '03*, Washington, DC, USA, August 24–27, pp. 256–265. ACM Press, New York City, NY, USA.

[20] Hilderman, R. and Peckham, T. (2007) Statistical Methodologies for Mining Potentially Interesting Contrast Sets. In Guillet, F. and Hamilton, H. (eds), *Quality Measures in Data Mining*. Springer, Berlin.

[21] Kralj Novak, P., Lavrač, N. and Webb, G. (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, **10**, 377–403.

[22] Azevedo, P. (2010) Rules for contrast sets. *Intell. Data Anal.*, **14**, 623–640.

[23] del Jesus, M., Gonzalez, P., Herrera, F. and Mesonero, M. (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. *Trans. Fuzzy Syst.*, **15**, 578–592.

[24] May, M. and Ragia, L. (2002) Spatial Subgroup Discovery Applied to the Analysis of Vegetation Data. In Karagiannis, D. and Reimer, U. (eds), *Practical Aspects of Knowledge Management*. Springer, Berlin.

[25] Allison, D., Cui, X., Page, G. and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Revi. Genet.*, **5**, 55–65.

[26] Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci.*, **102**, 15545–15550.

[27] Kim, S.-Y. and Volsky, D. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinform.*, **6**, 144.

[28] Cui, X. and Churchill, G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Gen. Biol.*, **4**, 210.1–210.10.

[29] Baldi, P. and Long, A. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

[30] Mootha, V. *et al.* (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

[31] Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

[32] Aoki-Kinoshita, K. and Kanehisa, M. (2007) Gene Annotation and Pathway Mapping in KEGG. In Walker, J. and Bergman, N.H. (eds), *Comparative Genomics*. Humana Press, New York City, NY, USA.

[33] Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L., Rhee, S. and Stitt, M. (2004) MapMan: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.

[34] Srikant, R. and Agrawal, R. (1995) Mining Generalized Association Rules. *Proc. VLDB '95*, Zurich, Switzerland, September 11–15, pp. 407–419. Morgan Kaufmann Publishers, San Francisco, CA, USA.

[35] The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

[36] Podpečan, V. *et al.* (2011) SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinform.*, **12**, 416.

[37] Suzuki, E. and Tsumoto, S. (2000) Evaluating Hypothesis-Driven Exception-Rule Discovery with Medical Data Sets. *Proc. PADKK '00*, Kyoto, Japan, April 18–20, pp. 208–211. Springer, Berlin.

[38] Westfall, P. and Young, S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, Hoboken, NJ, USA.

[39] Bender, R. and Lange, S. (2001) Adjusting for multiple testing—when and how? *J. Clin. Epidemiol.*, **54**, 343–349.

[40] Ge, Y., Dudoit, S. and Speed, T. (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

[41] Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

[42] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodol.)*, **57**, 289–300.

[43] Vavpetič, A. and Lavrač, N. (2012) Semantic subgroup discovery systems and workflows in the SDM-toolkit. *Comput. J.*, advance access published online 4 June 2012.

APPENDIX. PERMUTATION TEST

Subgroup discovery methods typically evaluate a large number of potentially interesting subgroups. It is possible that some of them are apparently statistically significant just by chance. To address the multiple testing problem, that is, to control the type I error (false positive) rates, we perform a permutation test to obtain adjusted p-values (see, e.g. [38–40]). We randomly permute the classes (class attribute values) and calculate the p-value for each subgroup. We repeat this first step for 10 000 permutations, create a histogram by the p-values of each permutation's best subgroup and estimate the (corrected) p-value of the original subgroups using the histogram: The corrected p-value is the relative number of permutations, including the original one, in which the best p-value is smaller or equal to the original p-value. This approach returns only an approximation of the exact p-values, which is sufficient enough for our application, where we primarily use the resulting corrected p-values to rank the subgroups. For stronger statistical tests, one can use a method such as Holm's simple sequentially rejective multiple test procedure [41] or the false discovery rate [42] instead.