Contrast Mining from Interesting Subgroups

Laura Langohr[1], Vid Podpečan[2], Marko Petek[3],
Igor Mozetič[2], and Kristina Gruden[3]

[1] Department of Computer Science and
Helsinki Institute for Information Technology (HIIT),
University of Helsinki, Finland
laura.langohr@cs.helsinki.fi
[2] Department of Knowledge Technologies,
Jožef Stefan Institute, Ljubljana, Slovenia
{vid.podpecan,igor.mozetic}@ijs.si
[3] Department of Biotechnology and Systems Biology,
National Institute of Biology, Ljubljana, Slovenia
{marko.petek,kristina.gruden}@nib.si

Abstract. Subgroup discovery methods find interesting subsets of objects of a given class. We propose to extend subgroup discovery by a second subgroup discovery step to find interesting subgroups of objects specific for a class in one or more contrast classes. First, a subgroup discovery method is applied. Then, contrast classes of objects are defined by using set theoretic functions on the discovered subgroups of objects. Finally, subgroup discovery is performed to find interesting subgroups within the two contrast classes, pointing out differences between the characteristics of the two. This has various application areas, one being biology, where finding interesting subgroups has been addressed widely for gene-expression data. There, our method finds enriched gene sets which are common to samples in a class (e.g., differential expression in virus infected versus non-infected) and at the same time specific for one or more class attributes (e.g., time points or genotypes). We report on experimental results on a time-series data set for virus infected potato plants. The results present a comprehensive overview of potato's response to virus infection and reveal new research hypotheses for plant biologists.

1 Introduction

Subgroup discovery is a classical task in data mining for finding interesting subsets of objects. We extend subgroup discovery by a second subgroup discovery step to find interesting subgroups of objects of a specific class in one or more contrast classes. Contrast classes can represent, for example, different time points or genotypes. Their exact definition depends on the interest of the user. We build on a generic assumption that objects are grouped into classes and described by features (e.g., terms). Often several terms can be summarized under a more

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 390–406, 2012.

general term. We use hierarchies to incorporate such background knowledge about terms. We are not concerned whether objects represent individuals, genes, or something else, and neither what features, classes, and hierarchies represent. Consider the following examples.

In bioinformatics a common problem is that high-throughput techniques and simple statistical tests produce rankings of thousands of genes. Life-scientists have to choose few genes for further (often expensive and time consuming) experiments. Genes can be annotated, for example, by molecular functions or biological processes, which are organized as hierarchies. A life-scientist might be interested in studying an organism in virus infected and non-infected condition (classes) at different time points after infection (contrast classes). In this context, subgroup discovery is known as gene set enrichment, where genes represent features and the aim is to find subgroups of features. In contrast, to fit the retrieval of gene sets into the general subgroup discovery context, we consider genes as objects, their ranking values and their annotations as features. See Table 1 for a line-up of the terms used in the two communities.

We report on experimental results on a time-series data set for virus infected *Solanum tuberosum* (potato) plants. As *S. tuberosum* has only sparsely biological annotations, we use bisociations. Bisociations are concepts that are bridging two domains which are connected only very sparsely or not at all [1]. In our experiments we transfer knowledge from the well studied model plant *A. thaliana* to *S. tuberosum*, our plant under investigation.

Table 1. Synonyms from different communities

Subgroup Discovery	Bioinformatics
object or instance	gene
feature or attribute 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 genotype
class (or class attribute value), e.g., positive/negative	differential/non-differential gene expression
subgroup of objects	gene set
interesting subgroup	enriched gene set

In sociology objects are individuals which are described by different features. For example, bank customers can be described by their occupation, location, loan and insurance type. An economist then might be interested in comparing bank customers who are big spender (classes) and those who are not, before and after the financial crisis (contrast classes). Consider, as a toy example, bank customers in Table 2 and four background hierarchies in Fig. 1. The economist

might know that before the financial crisis there were more big spenders than afterwards. Other, perhaps less obvious subgroups, can be more interesting. For example, the economist might not expect that the subgroup described by the term Ljubljana is statistically significant for a contrast class "after financial crisis" in comparison to the contrast class "before the financial crisis".

While subgroup discovery has been addressed in different applications before (see Section 2 for related work). We propose and formulate the problem of subgroup discovery from interesting subgroups and describe how well-known algorithms can be combined to solve the problem (Section 3). In Section 4 we show how these definitions can be applied to find interesting subgroups of genes. We report on experimental results on a time-series data set for virus infected potato plants in Section 5. In Section 6 we conclude with some notes about the results and future work.

Table 2. Bank customers described by features: occupation (OCC), location (LOC), loan (LOAN), insurance (INS) (adapted from Kralj Novak et al. [2]). Different classes are big spender (BSP) as well as before/after financial crisis.

	Before financial crisis					After financial crisis				
ID	OCC	LOC	LOAN	INS	BSP	OCC	LOC	LOAN	INS	BSP
1	private	Maribor	flat	yes	yes	private	Maribor	flat	yes	yes
2	private	Piran	no	no	yes	private	Ljubljana	no	no	yes
3	private	Ljubljana	flat	no	yes	private	Ljubljana	no	no	yes
4	public	Ljubljana	flat	yes	yes	private	Ljubljana	no	no	yes
5	public	Maribor	no	yes	yes	private	Maribor	no	yes	yes
6	private	Maribor	no	no	yes	unemployed	Maribor	no	no	no
7	private	Ljubljana	car	no	yes	unemployed	Ljubljana	car	no	no
8	public	Maribor	no	no	yes	unemployed	Maribor	no	no	no
9	unemployed	Maribor	no	no	yes	unemployed	Ljubljana	no	no	no
10	private	Ljubljana	no	yes	no	private	Ljubljana	no	yes	no
11	private	Piran	no	no	no	unemployed	Piran	no	no	no
12	public	Piran	car	yes	no	public	Piran	car	yes	no
13	unemployed	Piran	no	no	no	unemployed	Piran	no	no	no
14	unemployed	Ljubljana	flat	no	no	unemployed	Ljubljana	no	no	no
15	unemployed	Piran	car	no	no	unemployed	Ljubljana	car	no	no

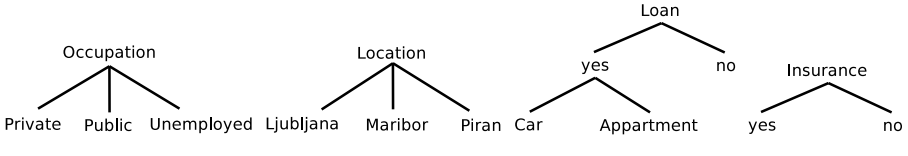

Fig. 1. Bank account feature ontologies (adapted from Kralj Novak et al. [2])

2 Related Work

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 described by rules (or logical implications) $Condition \mapsto Subgroup$, where the antecedent $Condition$ is a conjunction of attributes (e.g., terms) and the consequent $Subgroup$ is a set of objects.

Subgroup discovery methods find interesting subgroups of objects of a specific class compared to a complementary class. A subgroup of objects is interesting, when the feature values within the subgroup differ statistical significant from the feature values of the other objects. To analyze the constructed subgroups we use Fisher's exact test [5] and a simple test of significance. Alternatively, other statistical tests, like χ^2 test can be used.

Various application areas exist: sociology [6,7], marketing [8], vegetation data [9] or transcriptomics [10] amongst others. In sociology objects typically represent individuals and the aim is to find interesting subgroups of individuals.

In bioinformatics subgroup discovery is known as gene set enrichment. There, objects represent genes and the aim is to find subgroups of genes. A gene set is interesting (or enriched) if the differential expression of the genes of that gene set are statistically significant compared to the rest of the genes. The expression values of several samples are transformed into one feature value, called differential expression, and the genes are partitioned into two classes: differentially and not differentially expressed. Then, subgroup discovery methods find enriched gene sets. Alternatively, gene set enrichment analysis (GSEA) [11] or parametric analysis of gene set enrichment (PAGE) [12] can be used to analyze whether a subgroup is interesting (a gene set is enriched) or not. Both methods use not a partitioning of the genes into two classes, but a ranking of differential expressions instead.

Subgroup discovery differs from typical *time series analysis* where one observation per time point is given. Recently, different approaches have been described which split time series into shorter time-windows to be clustered in separated groups [13] or to find interesting subgroups [14,15]. However, subgroup discovery is not restricted to time series. In addition to time points it can also compare other types of classes, for example, healthy individuals compared to virus infected ones.

Contrast set mining aims to understand the differences between contrasting groups [16]. It is a special case of rule discovery [17] that can be effectively solved by subgroup discovery [18]. It is thus a generalization of subgroup discovery, in which two contrast classes are defined, in contrast subgroup discovery, where one class and it's complement are used.

Association rules describe associations like Y tends to be in the database if X is in it, where X and Y are item sets (sets of terms) [19]. *Exception rules* are association rules which differ from a highly frequent association rule [20]. Alike in our approach they aim to find unexpected rules. Their approach differs from the one presented here, as we are not only interested in finding subgroups

in one specific class, but in set theoretic combinations like intersections or set differences of subgroups found by a first subgroup discovery instance.

Frequent item set mining aims to find item sets describing a set of transactions (a subgroup) that are frequent [21]. Similar to the approach presented here, some methods intersect transactions to find closed frequent item sets [22,23,24].

Descriptive induction algorithms aim to discover individual rules defining interesting patterns in the data. This includes association rule learning [19], clausal discovery [25], contrast set mining [16] and subgroup discovery [6,7] amongst others. In contrast, *predictive induction* aims to construct rules to be used for classification and/or prediction [4]. We will focus on descriptive induction, even though our proposed approach could be adapted for predictive induction.

Semantic data mining denotes data mining methods which use background knowledge to improve pattern discovery and interpretation by using semantic annotations of objects as features [2]. Michalski [4] describes different types of background knowledge which can be subsumed under the term *ontology*. An ontology is a representation of a conceptualization and is often represented by a hierarchy, where nodes represent concepts and edges a subsumption relations [26]. Several ontologies can be modeled by a single ontology [27].

In biology commonly used ontologies include Gene Ontology (GO)[1] [28] and Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO)[2] [29]. GoMapMan[3] is an extension of the MapMan [30] ontology for plants used in our experiments. These ontologies are hierarchical vocabularies of gene annotations (semantic descriptors) organized as a directed acyclic graphs. Nodes represent molecular functions, biological processes or cellular components in GO, molecular pathways in KEGG and plant's molecular functions or biological processes in GoMapMan. Edges represent "is a" or "part of" relationships between the concepts (nodes).

Ontologies are extensively used in gene set enrichment [11,12]. Other application areas include association rule mining [27,31], where the transactions are either extended [27] or frequent item sets are generated one level at at a time [31]. Here, we use the subgroup construction method by Trajkovski et al. [32], which combines terms from the same level as well as from different levels.

3 Contrast Mining from Interesting Subgroups

Given a set of objects described by features and different classes of objects, the goal is to find interesting subgroups of objects of a specific class in one or more contrast classes. That is, for example, to find interesting subgroups specific for big spenders (class) after the financial crisis (contrast class).

Our approach finds such subgroups by dividing the task into three steps: First, interesting subgroups are found by a subgroup discovery method. Second, contrast classes on those subgroups are defined by set theoretic functions. Third,

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

subgroup discovery finds interesting subgroups in the contrast classes. Next, we will describe each step in detail.

3.1 Subgroup Discovery (Step 1)

To find interesting subgroups, we use search for enriched gene set (SEGS) [32], a method developed for gene set enrichment analysis, but not restricted to this application area [2].

First, all subgroups that contain at least a minimal number of objects are constructed by a depth-first traversal [32]. Afterwards, the constructed subgroups are analyzed if they are statistically significant for the class of interest.

Construction of Subgroups. We use hierarchies of terms as background knowledge to construct subgroups that contain at least a minimal number of objects. Subgroups are constructed by individual terms and logical conjunctions of terms.

Subgroup Construction by Individual Terms. Let S be the set of all objects and T the union of all terms of n background knowledges. Each term $t \in T$ defines a subgroup $S_t \subset S$ that consists of all objects s where feature value t is true, that is, are annotated by term t:

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

Subgroup Construction with Logical Conjunctions. Subgroups can be constructed by intersections, which are described by logical conjunctions of terms. Let $S_1, \ldots,$ S_k be k subgroups described by terms $t_1, \ldots, t_k$. Then, the logical conjunction of k terms defines the intersection of k subgroups:

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

Example 1. In Table 2, before the financial crisis, the conjunction *Ljubljana* $\wedge$ $\neg$*Insurance* defines a subgroup of three bank customers $\{3, 7, 14\}$.

A subgroup description can be seen as the condition part of a rule *Condition* $\mapsto$ *Subgroup* [33]. If an object is annotated by several terms, it is a member of several subgroups. A subgroup might be a subset of another subgroup. In particular, consider the example hierarchies in Figure 1. Then, an object that is annotated by a term t is also annotated by its ancestors.

To construct all possible subgroups one ontology is used, where the root has n children, one for each 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 of objects. Therefore, we extend a 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. Furthermore, in each recursion we add another term to the rule to obtain intersections of two or more subgroups and test if the intersection represents at least a minimal number of objects.

Analysis of Constructed Subgroups. Statistical tests can be used to analyze if the constructed subgroups are interesting, that is, the feature values within the subgroup differ statistically significant from the feature values of the other objects with respect to given classes A and B. For each subgroup $S_t \subset S$ the data is arranged in a table:

	A	B
S_t	n_{11}	n_{12}
$S \setminus S_t$	n_{21}	n_{22}

where $n = |S| = n_{11} + n_{12} + n_{21} + n_{22}$, n_{11} is the number of objects in S_t that are annotated by A, n_{12} is the number of objects in S_t that are annotated by B, n_{21} is the number of objects in $S \setminus S_t$ that are annotated by A, and n_{21} is the number of objects in $S \setminus S_t$ that are annotated by B.

Fisher's Exact Test. Fisher's exact test evaluates if the equal proportions and the observed difference is within what is expected by chance alone or not [5]. The probability of observing each possible table configuration is calculated by

$$P(X = n_{11}) = \binom{n_{11}+n_{12}}{n_{11}} \binom{n_{21}+n_{22}}{n_{21}} / \binom{n}{n_{11}+n_{21}} . \tag{3}$$

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

$$p = \sum_{i=0}^{n_{11}} P(X = i) . \tag{4}$$

Example 2. Consider the bank customers in Table 2, the condition *Maribor* and the class big spender versus not big spender and a significance level α. There are five bank customers in Maribor: $S_t = \{1, 5, 6, 8, 9\}$, which are all big spenders. Hence, the p-value is $p \approx 0.043956$.

Test of Significance. To address the multiple testing problem, that is, that subgroups might have occurred by chance alone, we correct the p-values. Therefore, we randomly permute the genes and calculate the p-value for each subgroup. We repeat this first step for $1,000$ permutations, create a histogram by the p-values of each permutation's best subgroup, and estimate the (corrected) p-value using the histogram: The corrected p-value is the reciprocal of the permutations in which the p-value obtained by Fisher's exact test is smaller than all p-values obtained from the permutations. For example, if the p-value obtained by Fisher's exact test is in all permutations smaller, then the corrected p-value is $p = 0.001$. If the corrected p-value is smaller than the given significance level α then the feature values within the subgroup differ statistical significantly from the feature values from the other objects and we call the subgroup interesting and the subgroup is called interesting.

3.2 Construction of Contrast Classes (Step 2)

Let $S_1, \ldots, S_n$ denote the interesting subgroups found for n classes. Then, two contrast classes S_f and S_g are defined by two set theoretic functions f and g:

$$f(S_1, \ldots, S_n) = S_f \subseteq \bigcup_i S_i \ . \tag{5}$$

and $g(S_1, \ldots, S_n)$ is defined as the complement. If $g(\cdot)$ is defined as something else than the complement, the next step is contrast set mining rather than subgroup discovery (see [33] for a line-up of both approaches).

Which set theoretic functions should be used depends on the objective. For example, if we aim to find interesting subgroups which are common to all classes, then $f(\cdot)$ is defined as the set of objects occurring in at least one interesting subgroup of each class:

$$f(S_1, \ldots, S_n) = \bigcap_{i \in \{1,\ldots,n\}} S_i \ . \tag{6}$$

Hence, every object of S_f occurred in each class in at least one interesting subgroup.

Alternatively, if the aim is to find interesting subgroups which are specific for class k, then $f(\cdot)$ is defined as the set of objects only occurring in interesting subgroups found for k^{th} class:

$$f(S_1, \ldots, S_n) = S_k \setminus \bigcup_{\substack{i \in \{1,\ldots,n\}, \\ i \neq k}} S_i \ . \tag{7}$$

Hence, every object in S_f occurred in one or more interesting subgroups of class S_k, but not in a single one of the other classes.

Example 3. Consider again the bank customers in Table 2, subgroups with at least four bank customers and $\alpha = 0.3$ (for sake of simplicity we consider a relatively high significance level in this toy example). For the "before financial crisis" class we obtain four subgroups: *Maribor*, *Maribor* $\wedge$ $\neg$*Loan*, *Piran*, and *Unemployed*. The set of bank customers described by at least one of them is $S_1 = \{1, 5, 6, 8, 11, \ldots, 15\}$. For the "after financial crisis" class we obtain two subgroups: *Private* and *Unemployed* and the set $S_2 = \{1, 2, 3, 6, \ldots, 11, 13, 14, 15\}$. Then the sets $S_f = S_2 \setminus S_1 = \{2, 3, 7, 10\}$ and $S_g = S_1$ specify contrast classes.

3.3 Subgroup Discovery (Step 3)

We find interesting subgroups in contrasting classes by a second subgroup discovery instance, where the two classes are now the sets S_f and S_g. The p-values are calculated by (3) and (4), followed by a test of significance.

Example 4. Given the contrast classes (sets) of bank customers $S_f = \{2, 3, 7, 10\}$ and $S_g = \{1, 5, 6, 8, 9, 11, \ldots, 15\}$ we analyze the statistical significance of subgroups with respect to these contrast classes. The condition *Ljubljana* has after

the financial crisis eight bank customers $\{2, 3, 4, 7, 9, 10, 14, 15\}$, from which four are in S_f and three in S_g. Hence, we obtain a p-value of $p = 0.0699301$. Next, we test the p-value for significance to assure we did not obtain the subgroup by chance alone. In the first subgroup discovery instance, we did not not obtain *Ljubljana* as logical condition. When compared to the contrast class "before financial crisis", and assuming it passed the significance test, *Ljubljana* is found to be statistically significant for the contrast class "after financial crisis".

4 An Instance of Our Method: Gene Set Enrichment from Enriched Gene Sets

Next, we will discuss how our proposed method can be applied in the area of gene set enrichment. In gene-expression experiments objects are genes and features are their annotations by, for example, GO and KEGG terms. Here, our aim is to find enriched gene sets of a specific class (e.g., virus infected plants) in one or more other classes (e.g., different time points). Next, we describe measures used for transforming the expression values of several samples (e.g., different individuals). into a feature value, called differential expression, and how the constructed gene sets are analyzed for statistical significance.

Measures for Differential Expression. After preprocessing the data (including microarray image analysis and normalization) the genes can be ranked according to their gene expression.

Fold change (FC) is a metric for comparing the expression level of a gene g between two distinct experimental conditions (classes) A and B [10]. It is the log ratio of the average gene-expression levels with respect to two conditions [34]. However, FC values do not indicate the level of confidence in the designation of genes as differently expressed or not.

The *t-test statistic* is a statistical test to determine the statistically significant difference of gene g between two classes A and B [10]. Though, the probability that a real effect can be identified by the t-test is low if the sample size is small [34]. A Bayesian t-test is advantageous if few (that is, two or three) replicates are used only, but no advantage is gained if more replicated are used [35]. In our experiments we used four replicates and therefore will use the simple t-test.

Analysis of Gene Set's Enrichment. For the enrichment analysis of gene sets statistical tests like Fisher's exact test [5] can be used. Alternatively, GSEA and PAGE can be used. We next describe each of them.

Fisher's Exact Test. In the gene set enrichment setting S_t is the gene set analyzed and $S \setminus S_t$ is the gene set consisting of all other genes. The two classes are differential expression and non-differential expression. To divide the genes into two classes 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 tested for significance.

Gene Set Enrichment Analysis (GSEA) [11]. Given a list $L = \{g_1, \ldots, g_n\}$ of n ranked genes, their expression levels $e_1, \ldots, e_n$, and a gene set S_t , GSEA evaluates whether S_t's objects are randomly distributed throughout L or primarily found at the top or bottom [36]. 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:

$$ES(S_t) = \max_{i \in \{1,\ldots,n\}} \left| \sum_{\substack{g_j \in S_t \\ j \leq i}} \frac{|e_j|^p}{n_w} - \sum_{\substack{g_j \notin S_t \\ j \leq i}} \frac{1}{n - n_w} \right| \tag{8}$$

where $n_w = \sum_{g_j \in S_t} |e_j|^p$. If the enrichment score 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 end of the list L. The exponent p controls the weight of each step. $ES(S_t)$ reduces to the standard Kolmogorov-Smirnov statistic if $p = 0$:

$$ES(S) = \max_{i \in \{1,\ldots,n\}} \left| \sum_{\substack{g_j \in S_t \\ j \leq i}} \frac{1}{|S_t|} - \sum_{\substack{g_j \notin S_t \\ j \leq i}} \frac{1}{|S| - |S_t|} \right|. \tag{9}$$

The significance of $ES(S_t)$ is then estimated by permutating the sample labels, reordering the genes, and re-computing $ES(S_t)$. From 1,000 permutations a histogram is created and the nominal p-value for S_t is estimated by using the positive (or the negative) portion if $ES(S_t) > 0$ (or $ES(S_t) < 0$, respectively).

Parametric Analysis of Gene Set Enrichment (PAGE). PAGE is a gene set enrichment analysis method based on a parametric statistical analysis model [12]. For each gene set S_t a Z-score is calculated, which is the fraction of mean deviation to the standard deviation of the ranking score values:

$$Z(S_t) = (\mu_{S_t} - \mu) \frac{1}{\sigma} \sqrt{|S_t|} \tag{10}$$

where σ is the standard deviation and μ and μ_{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 test of significance.

Using normal distributions for statistical inference makes PAGE computationally lighter than GSEA which requires permutations. On the other hand, GSEA makes no assumptions about the variability and can be used if the distribution is not normal or unknown. Kim and Volsky [12] studied different data sets for which PAGE generally detected a larger number of significant gene sets than GSEA. Trajkovski et al. [32] used the sum of GSEA's and PAGE's p-values,

weighted by percentages (e.g., one third of GSEA's and two third of PAGE's or half of both). Hence, gene sets with small p-values for GSEA and PAGE are output as enriched gene sets.

In the second gene set enrichment analysis instance, we want to analyze subgroups with respect to the constructed contrast classes, and not with respect to to the differential expression. Now, we have two classes, but not a ranking and thus GSEA and PAGE cannot be used for analyzing the constructed gene sets. Statistical test for categorical analysis can still be used. We use Fisher's exact test to compare the two classes S_f and S_g against each other.

5 Experiments

For our experiments we use a *Solanum tuberosum* (potato) time course gene-expression data set for virus infected and non-infected plants. The data set consists of three time points: one, three and six 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 which are common to virus infected samples compared to non-infected samples (classes in subgroup discovery of Step 1), and at the same time specific for one or all time points (classes in subgroup discovery of Step 3).

Test Setting. Recently, *S. tuberosum*'s genome has been completely sequenced [37], but only few GO or KEGG annotations of *S. tuberosum* genes exist. However, plenty GO and KEGG annotations exist for the well studied model plant *Arabidopsis thaliana*. We use homologs between *S. tuberosum* and *A. thaliana* to make gene set enrichment analysis for *S. tuberosum* possible. There are more than 26.000 homologs provided by the POCI consortium [38] for more than 42.000 *S. tuberosum* genes. We consider only the best (with respect to the e-value) in case there are several homologs. Gene set enrichment analysis 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*.

In parallel, we built potato ontologies independently using Blast2GO[4] to obtain homologue sequences in NCBI (BLASTX with high scoring segment pair (HSP) length 33 and e-value $1e-15$) and their GO annotations (GO weight 5, cutoff 55 and e-value $1e-15$). In this case, enrichment analysis is performed using the gene IDs and expression values of *S. tuberosum*, and the GO and KEGG annotations obtained with Blast2GO.

For both approaches we carried out gene set enrichment experiments in an Orange4WS[5] workflow [39]. We restricted gene sets to contain at minimum ten genes, the gene set description to contain at maximum four terms, and the p-value to be 0.05 or smaller. For analyzing the constructed gene sets in Step 1 we used Fisher's exact test, GSEA, PAGE and the combined GSEA and PAGE (equal percentages).

[4] `http://www.blast2go.org/`
[5] `http://orange4ws.ijs.si/`

We consider two types of contrast classes for gene set enrichment (Step 2): genes that are common to all classes compared to the genes occurring in some gene sets, but not in all (obtained by (6) and genes that are specific for one class compared to the genes of the gene sets of the other classes (obtained by (7). Fisher's exact test is used to analyze gene set enrichment in Step 3 for both approaches.

Results. Several subgroup descriptions that are known to relate to potato's response to virus infection were found. That is, our method reveals molecular functions, biological processes and pathways that have a central role in it. We are interested in assisting the biologist in generating new research hypotheses. Therefore, we evaluate our results by counting the number of gene set descriptions which were unexpected to a plant biologist to relate to potato's response to virus infection. 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 experimental conditions studied (here, related to the response of potato to viral infection).

The amount of enriched gene sets found for the *A. thaliana* homologs approach are shown in Table 3. and for the GO ontologies for potato genes approach in Table 4. For both approaches, both subgroup discoveries (Step 1 and 3) found few rules if any at all for the first and third day, whereas for the sixth day several rules are found. This matches well with the biological knowledge about potato's respond on virus infection: In the first days the potato activates the defense response, but the full effect can be witnessed only on day six.

The quantities of unexpected enriched gene sets found for the *A. thaliana* homologs approach are shown in Table 5. and for the GO ontologies for potato genes approach in Table 6. Few enriched gene sets are found in the first stage when using GSEA or the combination of GSEA and PAGE for analyzing the gene sets of the first stage. Hence, few enriched gene sets (if any at all) are found in Step 3. When using either Fisher's exact test or PAGE instead, more enriched gene sets are found, from which several are of interest to a plant biologist, suggesting one of these methods should be preferred.

The subgroups discovered in Step 3 revealed some enriched gene sets for the intersection, but none of them was more specific in comparison to the enriched gene sets found in Step 1 or even unexpected for the biologist. This is most likely due to the characteristic of a defense response: The gene expression of the first days (when activating the defense response) differs from the gene expression on day six (when the defense response is active) and therefore the intersection reveals only few enriched gene sets that are active at all time points.

For the set differences we obtain new and more specific gene sets. Some of them we did not find in the first stage, some other are more specific than in Step 1, both of interest for biologists. Hence, this shows that our proposed method reveals new enriched gene sets if the set theoretic functions are selected appropriately for the experiment and user's objective.

Table 3. Quantities of enriched gene sets found for the *A. thaliana* homologs approach for Fisher (F), GSEA (G), PAGE (P), and the combined approach of GSEA and PAGE with equal percentages (G+P)

		F	G	P	G+P
Step 1	first day	6	4	5	1
	third day	7	4	16	5
	sixth day	14	5	12	5
Step 3	first day set difference	9	0	7	1
	third day set difference	7	0	7	6
	sixth day set difference	21	5	16	6
	intersection	4	4	16	4

Table 4. Quantities of enriched gene sets found for the GO ontologies for potato genes approach for Fisher (F), GSEA (G), PAGE (P), and the combined approach of GSEA and PAGE with equal percentages (G+P)

		F	G	P	G+P
Step 1	first day	1	0	4	0
	third day	1	1	5	0
	sixth day	25	21	33	16
Step 3	first day set difference	15	0	7	0
	third day set difference	5	1	10	0
	sixth day set difference	42	2	34	3
	intersection	0	0	1	0

Table 5. Quantities of unexpected enriched gene sets found for the *A. thaliana* homologs approach for Fisher (F), GSEA (G), PAGE (P), and the combined approach of GSEA and PAGE with equal percentages (G+P). In Step 3 only unexpected enriched gene sets are counted which were new or more specific in comparison to Step 1.

		F	G	P	G+P
Step 1	first day	2	2	0	0
	third day	4	2	4	4
	sixth day	14	5	12	5
Step 3	first day set difference	1	0	4	0
	third day set difference	1	0	1	0
	sixth day set difference	11	1	4	1
	intersection	0	0	0	0

Table 6. Quantities of unexpected enriched gene sets found for the GO ontologies approach for Fisher (F), GSEA (G), PAGE (P), and the combined approach of GSEA and PAGE with equal percentages (G+P). In Step 3 only unexpected enriched gene sets are counted which were new or more specific in comparison to Step 1.

		F	G	P	G+P
Step 1	first day	0	0	1	0
	third day	1	1	2	0
	sixth day	24	21	28	16
Step 3	first day set difference	4	0	0	0
	third day set difference	0	0	2	0
	sixth day set difference	15	0	13	0
	intersection	0	0	0	0

6 Conclusion

We addressed the problem of subgroup discovery from interesting subgroups. After reviewing subgroup discovery we introduced the construction of contrast classes on the discovered subgroups. Subgroup discovery then finds interesting subgroups in those contrast classes. Thereby, we allow the user to specify contrast classes she is interested in, for example, she can choose to contrast several time points.

We showed how our approach works on an example of bank customers and applied it to a gene set enrichment application, a time-series data set for virus infected potato plants. The results indicate that our proposed approach reveals new research hypotheses for biologists.

Further experimental evaluation is planned, including experiments on other data sets and with more complex set theoretic functions. A careful interpretation of our results is needed as the subgroup discovery of the first step reduced the number of genes (objects) and hence Fisher was applied (in the third step) on a relatively small number of genes. Furthermore, gene set descriptions were often biologically redundant which we will address in future, for example, by clustering or filtering the obtained gene sets.

We will carry out a more extensive evaluation by analyzing the quality of gene sets descriptions which are unknown to relate to potato's virus response and visualize the gene sets and their relations with the enrichment map tool. We will evaluate quantity and quality of the genes of the unknown gene sets with Biomine, a search engine for visualization and discovery of non-trivial connections between biological entities, such as genes. Finally, some genes will be selected for wet-lab experiments, which may further the understanding of the biological mechanisms of virus response, particularly that of potatoes.

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

Nada Lavrač and Hannu Toivonen for their valuable comments and suggestions on the chapter.

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. Berthold, M.R. (ed.): Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250. Springer, Heidelberg (2012)
2. Kralj Novak, P., Vavpetič, A., Trajkovski, I., Lavrač, N.: Towards Semantic Data Mining with g-SEGS. In: SiKDD 2010 (2010)
3. Bruner, J., Goodnow, J., Austin, G.: A Study of Thinking. Wiley (1956)
4. Michalski, R.: A Theory and Methodology of Inductive Learning. Artificial Intelligence 20(2), 111–161 (1983)
5. van Belle, G., Fisher, L., Heagerty, P., Lumley, T.: Biostatistics: A Methodology for the Health Sciences, 2nd edn. Wiley series in probability and statistics. Wiley-Interscience (1993)
6. Klösgen, W.: Explora: a Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI (1996)
7. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Komorowski, J., Żytkow, J.M. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997)
8. del Jesus, M., Gonzalez, P., Herrera, F., Mesonero, M.: Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A Case Study in Marketing. Transactions on Fuzzy Systems 15, 578–592 (2007)
9. May, M., Ragia, L.: Spatial Subgroup Discovery Applied to the Analysis of Vegetation Data. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2002. LNCS (LNAI), vol. 2569, pp. 49–61. Springer, Heidelberg (2002)
10. Allison, D., Cui, X., Page, G., Sabripour, M.: Microarray Data Analysis: from Disarray to Consolidation and Consensus. Nature Reviews, Genetics 5, 55–65 (2006)
11. Mootha, V., Lindgren, C., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M., Patterson, N., Mesirov, J., Golub, T., Tamayo, P., Spiegelman, B., Lander, E., Hirschhorn, J., Altshuler, D., Groop, L.: PGC-1α-responsive Genes Involved in Oxidative Phosphorylation are Coordinately Downregulated in Human Diabetes. Nature Genetics 34(3), 267–273 (2003)
12. Kim, S.Y., Volsky, D.: PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 6(1), 144 (2005)

13. Antoniotti, M., Ramakrishnan, N., Mishra, B.: GOALIE, A Common Lisp Application to Discover Kripke Models: Redescribing Biological Processes from Time-Course Data. In: ILC 2005 (2005)
14. Antoniotti, M., Carreras, M., Farinaccio, A., Mauri, G., Merico, D., Zoppis, I.: An Application of Kernel Methods to Gene Cluster Temporal Meta-Analysis. Computers & Operations Research 37(8), 1361–1368 (2010)
15. Zoppis, I., Merico, D., Antoniotti, M., Mishra, B., Mauri, G.: Discovering Relations Among GO-Annotated Clusters by Graph Kernel Methods. In: Măndoiu, I.I., Zelikovsky, A. (eds.) ISBRA 2007. LNCS (LNBI), vol. 4463, pp. 158–169. Springer, Heidelberg (2007)
16. Bay, S., Pazzani, M.: Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery 5, 213–246 (2001)
17. Webb, G., Butler, S., Newlands, D.: On Detecting Differences between Groups. In: KDD 2003, pp. 256–265. ACM (2003)
18. Kralj Novak, P., Lavrač, N., Gamberger, D., Krstacic, A.: CSM-SD: Methodology for Contrast Set Mining through Subgroup Discovery. Journal of Biomedical Informatics 42(1), 113–122 (2009)
19. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.: Fast Discovery of Association Rules. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 307–328. AAAI (1996)
20. Suzuki, E.: Autonomous Discovery of Reliable Exception Rules. In: KDD 1997 (1997)
21. Agrawal, R., Imieliński, T., Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. In: SIGMOD 1993, pp. 207–216. ACM (1993)
22. Mielikäinen, T.: Intersecting Data to Closed Sets with Constraints. In: FIMI 2003 (2003)
23. Pan, F., Cong, G., Tung, A., Yang, J., Zaki, M.: Carpenter: Finding Closed Patterns in Long Biological Datasets. In: KDD 2003, pp. 637–642. ACM (2003)
24. Borgelt, C., Yang, X., Nogales-Cadenas, R., Carmona-Saez, P., Pascual-Montano, A.: Finding Closed Frequent Item Sets by Intersecting Transactions. In: EDBT/ICDT 2011, pp. 367–376. ACM (2011)
25. De Raedt, L., Dehaspe, L.: Clausal Discovery. Machine Learning 26, 99–146 (1997)
26. Gruber, T.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43, 907–928 (1995)
27. Srikant, R., Agrawal, R.: Mining Generalized Association Rules. In: VLDB 1995, pp. 407–419 (1995)
28. Khatri, P., Drăghici, S.: Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. Bioinformatics 21(18), 3587–3595 (2005)
29. Aoki-Kinoshita, K., Kanehisa, M.: Gene Annotation and Pathway Mapping in KEGG. In: Walker, J.M., Bergman, N.H. (eds.) Comparative Genomics, vol. 396, pp. 71–91. Humana Press (2007)
30. Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L., Rhee, S., Stitt, M.: MapMan: a User-driven Tool to Display Genomics Data Sets Onto Diagrams of Metabolic Pathways and Other Biological Processes. The Plant Journal 37(6), 914–939 (2004)
31. Han, J., Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases. In: VLDB 1995, pp. 420–431. Morgan Kaufmann Publishers Inc. (1995)
32. Trajkovski, I., Lavrač, N., Tolar, J.: SEGS: Search for enriched gene sets in microarray data. Journal of Biomedical Informatics 41(4), 588–601 (2008)

33. Kralj Novak, P., Lavrač, N., Webb, G.: Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research 10, 377–403 (2009)
34. Cui, X., Churchill, G.: Statistical Tests for Differential Expression in cDNA Microarray Experiments. Genome Biology 4(4), 210.1–210.10 (2003)
35. Baldi, P., Long, A.: A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-test and Statistical Inferences of Gene Changes. Bioinformatics 17(6), 509–519 (2001)
36. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., Mesirov, J.: Gene Set Enrichment Analysis: A Knowledge-based Approach for Interpreting Genome-wide Expression Profiles. PNAS 102(43), 15545–15550 (2005)
37. The Potato Genome Sequencing Consortium: Genome sequence and analysis of the tuber crop potato. Nature 475, 189–195 (2011)
38. Bioinformatics @ IPK Gatersleben: BLASTX against Arabidopsis, `http://pgrc-35.ipk-gatersleben.de/pls/htmldb_pgrc/f?p=194:5:941167238168085::NO` (visited on March 2011)
39. Podpečan, V., Lavrač, N., Mozetič, I., Kralj Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K.: SegMine Workflows for Semantic Microarray Data Analysis in Orange4WS. BMC Bioinformatics 12, 416 (2011)