J Intell Inf Syst
DOI 10.1007/s10844-013-0292-1

Semantic subgroup explanations

Anže Vavpetič · Vid Podpečan · Nada Lavrač

Received: 8 May 2013 / Revised: 20 September 2013 / Accepted: 19 November 2013

Abstract Subgroup discovery (SD) methods can be used to find interesting subsets of objects of a given class. While subgroup describing rules are themselves good explanations of the subgroups, domain ontologies can provide additional descriptions to data and alternative explanations of the constructed rules. Such explanations in terms of higher level ontology concepts have the potential of providing new insights into the domain of investigation. We show that this additional explanatory power can be ensured by using recently developed semantic SD methods. We present a new approach to explaining subgroups through ontologies and demonstrate its utility on a motivational use case and on a gene expression profiling use case where groups of patients, identified through SD in terms of gene expression, are further explained through concepts from the Gene Ontology and KEGG orthology. We qualitatively compare the methodology with the supporting factors technique for characterizing subgroups. The developed tools are implemented within a new browser-based data mining platform ClowdFlows.

Keywords Data mining · Semantic data mining · Subgroup discovery · Ontologies · Microarray data

A. Vavpetič (✉) · V. Podpečan · N. Lavrač
Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
e-mail: anze.vavpetic@ijs.si

V. Podpečan
e-mail: vid.podpecan@ijs.si

N. Lavrač
e-mail: nada.lavrac@ijs.si

A. Vavpetič · N. Lavrač
Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

N. Lavrač
University of Nova Gorica, Nova Gorica, Slovenia

Published online: 06 December 2013

1 Introduction

The paper first addresses the task of subgroup discovery, initially defined by Klösgen (1996) and Wrobel (1997), which is based both on classification and association discovery approaches. The goal is to find subgroups of individuals that are statistically important according to some property of interest of a given population of individuals. For example, a subgroup should be as large as possible and exhibit the most unusual distribution of the target class compared to the rest of the population.

Subgroup discovery methods can be used to find descriptions of objects of a given class—in binary as well as in multi-class problems. Subgroup descriptions, formed as rules with a class label in the rule conclusion and a conjunction of attribute values in the rule condition, typically provide sufficiently informative explanations of the discovered subgroups. However, with the expansion of the Semantic Web and the availability of numerous domain ontologies which provide domain background knowledge and semantic descriptors to the data, we are faced with the challenge of using this publicly available information also to provide explanations of rules initially discovered by standard symbolic data mining and machine learning algorithms. Approaches which would enhance symbolic rule learning with the capability of providing explanations of the rules also in terms of higher-level concepts than those used in rule descriptors, have a potential of providing new insights into the domain of investigation.

To give a simple example, suppose a standard subgroup discovery algorithm produces two rules for a dataset with patients (with the class *cancer*=0/1) and genes as attributes:

$$R_1 : (cancer = 1) \leftarrow (g_A = 1) \wedge (g_B = 1) \wedge (g_C = 0)$$
$$R_2 : (cancer = 1) \leftarrow (g_A = 0) \wedge (g_B = 1) \wedge (g_D = 1)$$

Each rule defines a subgroup of patients for which the right-hand side is true. These rules are by themselves explanatory in terms of single genes. But due to the existence of genetic regulatory networks, there are complex dependency structures between genes, e.g., multiple genes might be associated with a certain biological function. Using an ontology of biological knowledge (see next paragraph), we can find higher-level patterns on top of the gene-level patterns (such as rules R_1 and R_2). We propose that this can be achieved, for example, by taking R_1 and R_2 as the new classes and inducing new higher-level patterns by grouping the single genes into higher-level concepts defined by the ontology. An example higher-level rule E_1 (which we call an *explanation*) is:

$$E_1 : (cls = R_1) \leftarrow (c_1 = 1)$$

E_1 states that the patients defined by subgroup R_1 (the new target class) are characterized by the higher-level concept c_1 (e.g., a biological function), in contrast to patients from R_2. This is a higher-level statement, which takes into account multiple genes which are associated with the particular biological function c_1. This association knowledge is provided beforehand by the domain ontology.

We must emphasize that this explanatory step is not limited only to subgroup discovery. Essentially, the explanatory stage can be applied on any sets of examples that are of interest to the user, provided that a suitable ontology exists.

In this paper we show that such an additional explanatory step can be performed by using recently developed semantic subgroup discovery approaches (Podpečan et al. 2011a; Vavpetič and Lavrač 2013). The new methodology is show-cased on two use cases: a motivational use case of bank customers and on a gene expression profiling real-life use case.

The motivational use case showcases the methodology on a simple use case with banking customers and three simple ontologies, in order to illustrate the steps of the methodology.

In the gene expression use case, groups of patients of a selected grade of breast cancer, identified through subgroup discovery in terms of gene expression, are further explained through terms from the Gene Ontology[1] (GO) and Kyoto Encyclopedia of Genes and Genomes[2] (KEGG) and Entrez[3] gene-gene interaction data. The motivation for the use case in breast cancer patient analysis comes from the experts' assumption that there are several subtypes of breast cancer. Hence, in addition to distinguish between patients with breast cancer (the positive cases) and healthy patients, the challenge is first to identify breast cancer subtypes by finding subgroups of patients followed by inducing explanations in terms of identical biological functions, processes and pathways of genes, characterizing different molecular subtypes of breast cancer.

With the two use cases we demonstrate that the proposed approach is general and can be applied in any application area, provided the existence of domain specific ontologies.

The main contributions of the present work are as follows. First, inducing explanations of subgroups (or, e.g., clusters of instance), regardless of how the subgroups were detected, in terms of knowledge encoded in a domain ontology. Second, we have made our approach readily available on the web, as a reusable data mining workflow, which we hope will be a valuable resource for scientists, enabling them to use the workflow on new data, as well as adapt it for other use cases.

In addition, this work upgrades our early results (Vavpetič et al. 2012) in several ways. First, we have fully integrated our approach with the microarray analysis SegMine system (Podpečan et al. 2011a). Researchers using our tools can now also use the results of our methodology to query the Biomine search engine (Eronen and Toivonen 2012). Biomine essentially merges a large number of public biological databases into a common graph. The nodes in this graph are biological entities, while the edges are relations between them. Biomine offers advanced probabilistic graph search algorithms that can discover the parts of the graph most relevant to the given query. Examples of queries are: finding a neighborhood of a set of nodes or a graph connecting two sets of nodes. Biomine also offers a visualization tool for the user to explore the resulting subgraph.

Next, compared to our previous work where we made our tools available in the Orange4WS (Podpečan et al. 2011b) data mining platform, we have now moved to a new browser-based platform ClowdFlows (Kranjc et al. 2012). The main benefits of moving to ClowdFlows are: (a) no installation is required prior to using our tools (apart from an internet connection and a web browser), (b) scientific workflows and data can be shared by sharing a single URL, and (c) users can easily clone and adapt existing workflows to their own needs. We give an overview of the implementation, as well as discuss the pros and cons of the approach. In addition, the related work and the methodology are described in much more detail, enabling detailed methodology understanding and enabling its modification (upgrades by other researchers).

Additionally, the paper shows that the methodology is generally applicable for explaining groups of instances in any domain in which domain concepts are organized into ontologies and where data descriptions (attributes or attribute values) correspond to concepts from the ontologies. This is demonstrated with the two distinct use cases.

[1] http://www.geneontology.org/

[2] http://www.genome.jp/kegg/

[3] http://www.ncbi.nlm.nih.gov/sites/gquery

Finally, we compare our approach to the related *supporting factors* (Gamberger and Lavrač 2003) methodology, which is also used to characterize subgroups and can be of great help to the interpretation of subgroup discovery patterns of domain experts. The experiments show that supporting factors are more useful when concentrating on specific low-level attributes or features is desirable, but when more general descriptions are needed, they are not as easy to interpret as the method presented in this paper. This restriction is more apparent on gene expression data, since the supporting factors are given in terms of genes.

The paper is structured as follows. Section 2 discusses the related work. The proposed methodology is presented in detail in Section 3. The ClowdFlows platform and the implementation of the methodology are described in Section 4. In Section 5 the methodology is applied to two use case scenarios: a motivational use case and a use case with breast cancer gene expression (microarray). The presented methodology is also compared to the supporting factors methodology on the second use case. Section 6 concludes the paper and presents the plans for further work.

2 Related work

This section discusses the work related to the main steps of the proposed methodology. Given a complex multi-step approach, the related work covers subgroup discovery, contrast mining, subgroup explanation, and semantic data mining. Mining of enriched gene sets from gene expression data is also relevant for the biomedical use case presented in Section 5 (analysis of breast cancer data) which is used to evaluate the proposed methodology.

Subgroup discovery The problem of subgroup discovery was defined by Klösgen (1996) and Wrobel (1997) as search for population subgroups which are statistically interesting and which exhibit unusual distributional characteristics with respect to the property of interest. Subgroup descriptions are conjunctions of attributes and values which characterize the selected class of individuals. Several algorithms were developed for mining interesting subgroups using exhaustive search or using heuristic approaches: Explora (Klösgen 1996), APRIORI-C (Jovanoski and Lavrač 2001), APRIORI-SD (Kavšek and Lavrač 2006), SD-Map (Atzmüller and Puppe 2006), SD (Gamberger and Lavrač 2002), CN2-SD (Lavrač et al. 2004). These algorithms employ different heuristics to asses the interestingness of the discovered rules, which is usually defined in terms of rule unusualness and size.

Contrast mining Mining of contrasts in data has been recognized as one of the the fundamental tasks in data mining (Webb et al. 2003). The underlying idea is to discover and understand contrasts (differences) between objects of different classes, different time periods, spatial locations, objects within a class or various combinations of these. One of the first algorithms which has explicitly addressed the task of mining contrast sets is the STUCCO algorithm, developed by Bay and Pazzani (2001). It searches for conjunctions of attributes and values (contrast sets) which exhibit different levels of support in mutually exclusive groups, STUCCO enforces statistically sound results by employing testing for statistical significance and p-value correction along with minimum support threshold. Mining for contrasting sets is also related to exception rule mining as defined by Suzuki (1997, 2006) where the goal is to discover rare deviating patterns which complement strong base rules to form rule pairs. Suzuki (2006) defines an exception as something different from most of the rest of the data which can be also seen as a contrast to given data and/or existing

domain knowledge. A recent approach developed by Langohr et al. (2013), which proposes *contrasting subgroup discovery*, is closely related to the methodology presented in this paper. It extends classical subgroup discovery using a three-step approach and allows for the discovery of subgroups which cannot be found with classical subgroup discovery. Two subgroup discovery steps are complemented by the intermediate, contrast set definition step. In this intermediate step, the user explicitly defines two contrasting classes using set-theoretic functions and the subgroups discovered in the first step. In this way, generalized subgroups consisting of members from different classes can be discovered. While the approach is general and can be used on any data, it is especially well-suited for domains such as systems biology and biomedicine where comparing e.g., different time points in experimental data or several subtypes of a disease is a typical task.

Subgroup explanation The need of developing methods for presenting contrast sets to the user has already been recognized by Gamberger and Lavrač (2002) and Webb et al. (2003). Kralj Novak et al. (2009) have shown that contrast set mining, emerging pattern mining (Dong and Li 1999) as well as subgroup discovery can be viewed as variants of rule learning by providing appropriate definitions of compatibility; they also presented several subgroup visualization approaches, enabling subgroup comparison in terms of their size and distributional unusualness. However, to the best of our knowledge, neither different subgroup discovery algorithms nor the relatively efficient contrast/exceptional pattern mining algorithms like STUCCO (Bay and Pazzani 2001) and PEDRE (Suzuki 1997) address the representation and explanation of subgroups/contrasts using the available background knowledge and ontologies.

Semantic data mining While subgroup descriptions in the form of rules are relatively good descriptions of subgroups there is also abundance of background knowledge in the form of taxonomies and ontologies readily available to be incorporated to provide better high-level descriptions and explanations of discovered subgroups. Especially in the domain of systems biology the GO ontology, KEGG orthology and Entrez gene-gene interaction data are good examples of structured domain knowledge. The challenge of incorporating domain ontologies in data mining was addressed in the recent work on semantic data mining (SDM) (Hilario et al. 2011; Lavrač et al. 2011; Lawrynowicz and Potoniec 2011; Vavpetič and Lavrač 2013; Žáková et al. 2006).

Using a data mining ontology for meta-learning has been proposed in Hilario et al. (2011). In meta-learning the task is to use data mining techniques to improve base-level learning. The data mining ontology is used to (1) incorporate specialized knowledge of algorithms, data and workflows and to (2) structure the search space when searching for frequent patterns.

In Lawrynowicz and Potoniec (2011), they introduce an algorithm named Fr-ONT for frequent concept mining expressed in $\mathcal{EL}^{++}$ DL. In contrast to our work, the task they are solving is frequent concept mining and the hypothesis language they are using is $\mathcal{EL}^{++}$ description logic.

In Žáková et al. (2006) an engineering ontology of CAD (Computer-Aided Design) elements and structures is used as background knowledge to extract frequent product design patterns in CAD repositories and discovering predictive rules from CAD data.

This work is built upon the SDM toolkit developed by Vavpetič and Lavrač (2013). The toolkit includes two semantic data mining systems: SDM-SEGS and SDM-Aleph. SDM-SEGS is an extension of the earlier domain-specific algorithm SEGS (Trajkovski et al.

2008) which allows for semantic subgroup discovery in gene expression data. SEGS constructs gene sets as combinations of GO ontology terms, KEGG orthology terms, and terms describing gene-gene interactions obtained from the Entrez (Maglott et al. 2005) database. SDM-SEGS extends and generalizes this approach by allowing the user to input any set of ontologies in the OWL format and an empirical data collection which is annotated by domain ontology terms. SDM-SEGS employs ontologies to constrain and guide the top-down search of a hierarchically structured space of induced hypotheses. SDM-Aleph, which is built using the popular ILP system Aleph (Srinivasan 2007) does not have the limitations of SDM-SEGS, imposed by the domain-specific algorithm SEGS, and can accept any number of OWL ontologies as background knowledge which is then used in the learning process.

Semantic data mining and link discovery in enriched gene set analysis In the domain of systems biology, the SegMine methodology (Podpečan et al. 2011a) enables semantic analysis of microarray data by integrating the SEGS algorithm, GO and KEGG, and the Biomine system which integrates several public databases with a sophisticated algorithm for link discovery. Parts of the SegMine methodology can be reused in the methodology proposed in this paper for the specific use case of gene expression profiling. For example, link discovery can provide additional and potentially new information about the discovered important genes, subgroups and ontology terms.

Characterizing outliers In Angiulli et al. (2013), they consider a related task of characterizing attributes that account for a small group of anomalous examples-outliers. They define the notion of exceptional property and exceptionality score. They are designed to work especially with small samples. In contrast to our work, they focus mainly on small, anomalous groups of examples. The second main difference is that they do not try to generalize over the given attributes, since the exceptional properties are in terms of the original attributes.

Supporting factors The most relevant related work is the work by Gamberger and Lavrač (2003). In their work, they deal with characterizing subgroups through *supporting factors*. Supporting factors are features with significantly different value distributions that are not part of the subgroup description. Supporting factors are important, e.g., for medical decision making, which requires as much supportive evidence as possible. We compare our methodology with supporting factors in Section 5.2.

3 Methodology

Semantic subgroup discovery approaches such as SEGS, SDM-SEGS and SDM-Aleph can serve as explanatory subsystems in the presented methodology to semantically describe and explain contrasting groups in input data. This section presents the steps of the proposed methodology. The first step involves finding relevant sets of instances (relevant to the user) by applying a subgroup discovery algorithm, thus creating a new labeling for the instances in terms of their subgroup membership. The second step deals with ranking the attributes according to their ability to distinguish between the subgroups. The third step of the methodology induces symbolic explanations of a selected target set of instances (subgroup detected in the first step) by using ontological concepts.

We must emphasize again that the methodology consists of several steps, which are not novel by themselves, but are used in a novel fashion; also, each step of the components can be easily interchanged with several alternatives.

3.1 Identifying interesting sets of instances and creating a new labeling

To find a potentially interesting set of instances, the user can choose from a number of data mining algorithms. Data mining platforms such as Weka (Hall et al. 2009), Orange (Demšar et al. 2004), Orange4WS (Podpečan et al. 2009) and ClowdFlows (Kranjc et al. 2012) offer various clustering, classification and visualization techniques. A potentially interesting set of instances can be a cluster of instances, instances in a node of a decision tree, a set of instances revealed by a visualization method, a set of instances covered by a subgroup description, and others; in the following paragraphs we concentrate on subgroup discovery, but other techniques that define some sort of sets of examples can be used analogously (e.g., clustering; the user chooses between clusters instead of subgroups).

First, some basic notation needs to be established. Let $D = \{e_1, e_2, \ldots, e_n\}$ be a dataset of classified instances, called examples in the rest of this paper. Examples are defined by values of a set of attributes $A = \{a_1, a_2, \ldots, a_m\}$ and a continuous or discrete target variable y (note that unsupervised methods do not require a target variable). Let v_{ij} denote the value of attribute a_j for example e_i.

In the following, subgroups and clusters are represented as sets of examples. Let S_A and S_B denote two sets of examples ($S_A \cup S_B \subseteq D$) that are of interest to the user who wants to determine which groups of attributes (expressed as ontological concepts) differentiate S_A from S_B. Note that for subgroup descriptions the following must also hold: $S_A \cap S_B = \emptyset$, since it is typical that subgroups can overlap. This condition is of course not necessary for other settings like clustering.

Regardless of how S_A and S_B have been constructed, the new re-labeled dataset D' is formed as follows. The target variable y is replaced by a binary target variable y' and for each example e_i the new label c' is defined as:

$$c' = \begin{cases} 1, & \text{if } e_i \in S_A \\ 0, & \text{if } e_i \in S_B \end{cases}$$

Note that if D is unlabeled, the new target variable y' is added to the domain. We now illustrate how to determine S_A and S_B using a subgroup discovery (SD) approach.

SD algorithms induce symbolic subgroup descriptions of the form

$$(y = c) \leftarrow t_1 \wedge t_2 \wedge \ldots t_l$$

where t_j is a conjunct of the form $(a_i = v_{ij})$. If a_i is continuous and the selected subgroup discovery algorithm can deal with continuous attributes, t_i can also be defined as an interval such that $(a_i \geq v_{ij})$ or $(a_i \leq v_{ij})$. An example subgroup description constructed from the well known UCI lenses [4] dataset is:

$$(lenses = hard) \leftarrow (prescription = myope) \wedge$$
$$(astigmatic = yes) \wedge (tear\_rate = normal)$$

A subgroup description R can be also viewed as a set of constraints (conjuncts t_i) on the dataset, and the corresponding subgroup as a set of examples *cov(R)* which satisfy the constraints, i.e., examples covered by rule R.

[4] http://archive.ics.uci.edu/ml/datasets/Lenses

If the user is presented with a set of subgroup descriptions $R = \{R_1, R_2, \ldots, R_k\}$, then the set of examples S_A can be defined as $S_A = cov(R_i)$. For subgroup discovery S_B typically represents all other examples $S_B = D \setminus S_A$, because subgroups often overlap. For clustering S_B can be a single other cluster or a union of several clusters, depending on the user's preference.

To give a trivial example, suppose the subgroup discovery procedure returns three subgroup descriptions $R = \{R_1, R_2, R_3\}$ on the previously mentioned UCI lenses dataset. These are as follows:

$$R_1: \quad (lenses = hard) \leftarrow (age = young)$$
$$R_2: \quad (lenses = soft) \leftarrow (astigmatic = no)$$
$$R_3: \quad (lenses = none) \leftarrow (prescription = hypermetrope)$$

For example, R_1 covers all examples that have the attribute-value $age = young$; these examples constitute the rule's *coverage*.

The user can then select S_A and S_B, to give an example, as follows $S_A = cov(R_1)$ and $S_B = D \setminus S_A$. In this scenario S_A contains examples covered by R_1 and S_B contains all examples not covered by R_1.

3.2 Ranking of attributes

Once the re-labeled dataset D' is available, the attributes are assigned ranks according to their ability to distinguish between the two sets of examples S_A and S_B. The resulting ordered attributes and their scores will be used as input examples in the next step of the methodology. The generalizations of the attributes made via the ontological background knowledge will be the constituents of the resulting explanations.

To calculate the ranks, any attribute quality measure can be used, but in practice attribute ranking using the ReliefF (Robnik-Šikonja and Kononenko 2003) algorithm has proven to yield reliable scores for this methodology to work. In contrast to myopic measures (e.g., Gain Ratio), ReliefF takes into account the context of other attributes when evaluating an attribute. This is an important benefit when applying this methodology to datasets such as microarray data since it is known that there are dependencies among many genes.

The ReliefF algorithm works as follows. A random subset of examples of size $m \leq n$ is chosen. Each attribute starts with a ReliefF score of 0. For each randomly selected example e_i and each class c, k nearest examples are selected. The algorithm then goes through each attribute a_l and nearest neighbor e_j $(i \neq j)$, and updates the score of the attribute as follows:

- if e_i and e_j belong to the same class and at the same time have different values of a_l, then the attribute's score is decreased;
- if the examples have different attribute values and belong to different classes, then the attribute's score is increased.

This step of the methodology results in a list of ReliefF attribute scores $L = [(a_1, r_1), \ldots, (a_m, r_m)]$ where r_i is the ReliefF score representing the ability of attribute a_i to distinguish between sets S_A and S_B.

3.3 Inducing explanations using ontologies

At this stage of the methodology a semantic subgroup discovery algorithm (Lavrač et al. 2011; Vavpetič and Lavrač 2013) is applied to generate explanations using the list of ranked attributes L.

First, we need to formalize the notion of an ontology. An ontology is a conceptualization of a certain domain in terms of *concepts* and *relationships* between these concepts. An ontology is a directed acyclic graph, i.e., with no paths starting and ending on the same vertex, with concepts $C = \{c_1, c_2, \ldots c_n\}$ as vertices and relations $R = \{r_1, r_2, \ldots r_m\}$ as edges. Each relation is defined as a set of pairs of concepts: $r_i = \{(c_j, c_k)|c_j, c_k \in C\}$. Commonly used relationships are *subClassOf* (commonly referred to as *is-a*) and *partOf*. In this section we use the Gene Ontology as an example, which uses only these two relations.

Concepts and relations consistute the so-called *T-box* (the terminology). In order to connect the data (ranked attributes) to the ontology, we also require the *A-box* (the assertions). These we can view as a mapping $M = \{(a_i, c_j)|a_i \in A, c_j \in C\}$ of objects (in our case, attributes) onto concepts from the ontology. In the case of the Gene Ontology use case, the gene annotations represent our A-box, which defines which genes are annotated by which ontological concept (e.g., a biological function).

Each subgroup description (rule) induced by a semantic subgroup discovery algorithm represents one explanation, and each explanation is a conjunction of ontological concepts. The assumption here is that a domain ontology $O =< C, R >$ is available and that a mapping M between the attributes (or attribute values; we assume attributes in the rest of this section) and ontological concepts exists. For example, in the case of microarray data, an attribute (gene) IDH1 is mapped to (annotated by) the ontological concept *Isocitrate metabolic process* from the Gene Ontology, indicating that this gene takes part in this particular biological process. Thus, when translated into our methodology, each ontological concept, as well as each explanation, defines a set of attributes.

In other words, an existing semantic subgroup discovery algorithm is at this stage applied in a novel way - the algorithm internals are identical compared to when used for a standard subgroup discovery task.

Annotations enable the explanations to have strictly defined semantics, and from a data mining perspective, this information enables the algorithm to generalize better than by using attribute values alone. The explanations can be made even richer if additional relations among the attributes (or ontological concepts) are included in the explanations. Using the microarray example, genes are known to *interact*, and this information can be directly used to form explanations.

Currently, there are four publicly available SDM systems that can be used for the purpose of inducing explanations:

- SEGS (Trajkovski et al. 2008), a domain specific system for analyzing microarray data using the Gene Ontology, KEGG orthology, and Entrez gene-gene interactions,
- SDM-SEGS (Vavpetič and Lavrač 2013), the general purpose version of SEGS, that enables the use of OWL ontologies, but is limited to a maximum of four ontologies, i.e., the user needs to specify the rule language by defining up to four new roots of their ontology,
- SDM-Aleph (Vavpetič and Lavrač 2013), a general purpose SDM system based on the ILP system Aleph, that can use any number of OWL ontologies,
- Hedwig (Vavpetič et al. 2013), a new subgroup discovery SDM system, which builds upon the benefits of both SDM-SEGS and SDM-Aleph. Namely, it support the full RDFS ontology language and exploits the *subClassof* hierarchy to efficiently structure the search space.

All four systems focus on inducing explanations in the form of rules with conjuncts corresponding to ontological concepts. To illustrate how explanations are induced, consider that SEGS or SDM-SEGS (they have the same rule construction algorithm, but different

rule selection process) is selected to be used on a microarray domain. The algorithm used by SEGS and SDM-SEGS is the simplest of the four and is good for illustrating the semantic nature of the learning process, but it has its drawbacks. Namely, due to its simplicity only the *subClassOf* relation is exploited and one additional relation between the genes/attributes. SDM-Aleph is similar, except that it imposes no restrictions on the number of relations. On the other hand, Hedwig has no such limitations. The background knowledge can contain arbitrary relations, with *subClassOf* having a special status in that it is exploited to structure the search space.

Note that in the following description, genes can be thought of as instances or examples, since the algorithm is not limited only to genes. The idea behind SEGS as well as SDM-SEGS, illustrated on the problem of finding explanations for top-ranked genes, is as follows (Fig. 1 shows the rule construction algorithm in pseudo code).

The set of explanations/subgroup descriptions is constructed using top-down bounded exhaustive search according to the user-defined constraints (e.g., minimum support). The algorithm considers all explanations that can be formed by taking one concept from each ontology as a conjunct.

The input list L of ranked genes is first split into two classes. The set of genes above a selected threshold value is the set of differentially expressed genes for which a set of rules is constructed (these rules describe sets of genes which distinguish set S_A from set S_B).

The construction procedure starts with a default rule $top(X) \leftarrow$, with an empty set of conjuncts in the rule condition, which covers all the genes. With $top(X)$ we denote the target concept, which is in this case a set of attributes that near or at the *top* of the list L—thus good at distinguishing between the two sets. Next, the algorithm tries to conjunctively add the root concept of the first ontology (yielding e.g., $top(X) \leftarrow biological\_process(X)$) and if the new rule satisfies all of the size constraints (`MIN_SIZE` - minimum number

```
function nconstruct(rule, conj, k):
 # rule - the rule to specialize.
 # conj - the concept to add to the rule.
 # k - 'conj' is from the k-th ontology.

 # The set described by the current rule.
 newSet = intersect(set(rule), set(conj))

 # Is the set big enough?
 if newSet.size > MIN_SIZE:
   rule.add(conj)
   if 0 < rule.terms.size < MAX_TERMS:
     rules.add(rule)

   # Can the rule be extended?
   if rule.size < max(MAX_TERMS, MAX_ONT):
     construct(rule, ontologies[k+1], k+1)
     rule.remove(conj)

     # Extend the rule with all successors.
     for each child in children(conj):
       if set(child).size > MIN_SIZE:
         construct(rule, child, k)

 # Also check the interacting set.
 interactingSet = intersect(set(rule), interacts(set(conj)))
 if interactingSet.size > MIN_SIZE:
   rule.add('interacts(' conj ')')
   if rule.terms.size < MAX_TERMS:
     rules.add(rule)

 return rules
```

Fig. 1 Rule construction procedure of (SDM-)SEGS

of genes coverd by a rule, `MAX_TERMS` - maximum number of conjunctions in a single rule), it adds it to the rule set and recursively tries to add the root concept of the next ontology (e.g., $top(X) \leftarrow biological\_process(X) \wedge molecular\_function(X)$). In the next step all the child concepts of the current conjunct/concept are considered by recursively calling the procedure. Due to the transitivity of the *subClassOf* relation between concepts in the ontologies, the algorithm can employ an efficient pruning strategy. If the currently evaluated rule does not satisfy the size constraints, the algorithm can prune all rules which would be generated if this rule were further specialized.

Additionally, the user can specify gene interaction data by specifying the *interacts* relation. In this case, for each concept which the algorithm tries to conjunctively add to the rule, it also tries to add its interacting counterpart. For example, if the current rule is $top(X) \leftarrow c_1(X)$ and the algorithm tries to add the term/concept $c_2(X)$, then it also separately tries to append a compound term $interacts(X, Y) \wedge c_2(Y)$.

In SEGS, the constructed explanations are assigned scores using several established methods (e.g., GSEA Subramanian et al. 2005) and the significance of the explanations is evaluated using permutation testing (Trajkovski et al. 2008).

In our setting, the resulting descriptions correspond to subgroups of attributes (e.g., genes) which enable distinguishing between sets S_A and S_B. The interpretation is simple, due to the ontological concepts (conjuncts). Consider the following subgroup description:

$$top(X) \leftarrow immune\_system\_process(X) \wedge plasma\_membrane(X) \wedge interacts(X, Y) \wedge T\_cell\_receptor\_signaling\ pathway(Y).$$

This rule can be interpreted as follows. One of the top groups of genes (attributes) that are capable of distinguishing S_A from S_B, are the genes which take part in the *immune system process*, are part of the *plasma membrane* and interact with genes that are part of the *T cell receptor signaling pathway*.

4 Implementation

The described methodology was implemented in ClowdFlows (Kranjc et al. 2012), a publicly available workflow environment. We have extended the original implementation in the Orange4WS (Podpečan et al. 2011b) platform in order to make the experimental data and workflow, as well as the individual re-usable components easily accessible. As the ClowdFlows user interface runs entirely in a web browser there are no software requirements. Moreover, the developed workflows and the results of their execution can be shared by providing a link to the workflow. In the following we summarize the new implementation along with the most relevant features of ClowdFlows.

4.1 The ClowdFlows platform

ClowdFlows is a new generation platform for data mining which is implemented as a web application. It is based on the concept of *visual programming* which denotes the construction of complex procedures (workflows) from smaller building blocks (widgets) on a *canvas*. ClowdFlows offers a large collection of implemented algorithms, procedures and visualizations from different scientific fields: data mining, natural language processing, text mining, systems biology and inductive logic programming. New components can be implemented in the ClowdFlows server application or can be imported as web services. All included

components are available as widgets and can be used in the construction of data analysis workflows.

Two of the most important features of ClowdFlows are its graphical user interface and the database, which stores all information about components, workflows, data, and results. The graphical user interface, which runs as a web application, allows the user to interactively construct the workflow by placing the appropriate component on the canvas, set their paramenters, connects inputs and outputs and execute them. The database, on the other hand, stores all vital information and enables sharing of the constructed solutions, data, and experimental results by making the workflow accessible under a unique public URL. This greatly simplifies the evaluation of experimental results.

4.2 Implementation of the methodology workflow

The proposed methodology was implemented as a ClowdFlows workflow. Widgets from different ClowdFlows packages (such as utility widgets, e.g., *Load dataset*) as well as several newly developed components were deployed. First, the subgroup discovery package was used (some of these widgets are based on the Subgroup Discovery toolkit for Orange[5]). Second, the SDM-toolkit (Vavpetič and Lavrač 2013) and the SegMine tools (Podpečan et al. 2011a) from our previous work were also moved to ClowdFlows. Having these widgets made available within the platform, we were able to connect them into a workflow implementing our methodology. Figures 3 and 4 show two ClowdFlows workflows using our methodology for two use cases. Since the developed widgets are self-contained units with a well defined task, they can be re-used for other tasks as well (the roles of particular widgets are discussed in more detail Section 5).

5 Use cases

In this section we present the application of our methodology on two use cases. The first is a motivational use case intended to illustrate the methodology as well as showing how it can be applied using the ClowdFlows platform. The second use case is an application on real-world gene expression microarray data. On the second use case, we also apply the related supporting factors approach and qualitatively compare it to our approach.

5.1 Illustrative use case

This subsection further illustrates and motivates the use of the methodology on an easy-to-understand toy use case. First, we describe the dataset and cast the problem in our new framework. Next, we present the workflow developed for solving the toy problem by explaining each of the workflow's components.

This use case is an adaptation of the proof-of-concept semantic data mining dataset from Vavpetič and Lavrač (2013). Consider a bank which has the following data about its customers: place of living, employment, bank services used, which includes the account type, possible credits and insurance policies and so on. The attributes of the dataset are binary.

[5] http://kt.ijs.si/petra_kralj/SubgroupDiscovery/

Table 1 Table of bank customers described by several attributes and class 'big spender'

Doctor	Nurse	Munich	Rome	Classic	Gold	...	Big spender
1	0	0	0	1	0	...	yes
1	0	0	0	0	1	...	yes
0	0	1	0	0	1	...	yes
1	0	0	0	1	0	...	yes
0	0	0	0	0	1	...	yes
...	...	...	...	...	...	...	
0	0	0	0	0	1	...	no
0	1	0	0	1	0	...	no
0	0	0	0	1	0	...	no
0	0	0	0	0	1	...	no
0	0	0	0	1	0	...	no

For example, the attribute-value pair *Doctor=1* indicates that a particular customer is a doctor. The bank also labeled the clients as 'big spenders' or not and wants to find patterns describing big spenders. Table 1 presents a subset of the training data.

Suppose we also have three ontologies available as background knowledge for this problem: an ontology of banking services, an ontology of locations and an ontology of occupations, shown in Fig. 2. Note that the attributes of the dataset correspond to the leaves of the ontologies.

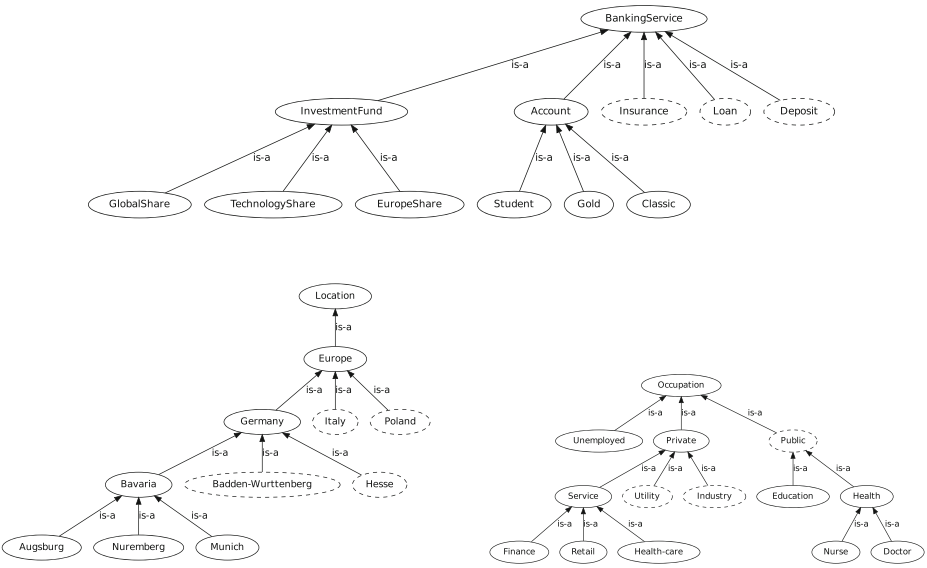

Fig. 2 The ontologies of banking services, locations and occupations. Concepts with omitted sub-concepts are drawn with a *dashed line*

In terms of our methodology, we first want to find descriptions of customers that are big spenders. After finding and selecting an interesting subgroup, we wish to use the knowledge of the domain ontologies to explain what are the differences between this subgroup of customers compared to the other customers.

Figure 3 shows a workflow developed to solve the described problem using our methodology. The workflow neatly follows the steps outlined in Section 3.

Step 1 *Identifying interesting sets of instances and creating a new labeling*, consists of the following components: the dataset is first uploaded (*Load dataset* widget), then standard subgroup discovery is run (*Build subgroups* widget) and the user is prompted to select one or more interesting subgroups (*Select subgroups* widget). The examples are then re-labeled, where the examples in the selected subgroup(s) represent one class, while the rest represents the other class (*Query data with subgroups* and *Table from sets of examples* widgets).

Step 2 *Ranking of attributes*, consists of a single *Ranker* widget, which uses the ReliefF algorithm to assign a score to each of the attributes and outputs a list of pairs (*attribute*, *score*).

Step 3 *Inducing explanations using ontologies*, is composed of one main widget: *SDM-Aleph*. This widget calls the SDM-Aleph web service, which employs the ontologies and the Aleph ILP system to produce subgroups. The widget accepts the list of ranked attributes, the OWL ontologies and the mapping between attribute names and ontology concepts (*Load mapping* and *Load ontology* widgets; note that these are actually *Load file to string* widgets renamed to reflect what they do). The *SDM-Aleph* widget returns a set of subgroups, which is displayed by the *Display subgroups* widget.

This public workflow contains an example experiment (using the dataset described above), where we have used the following settings. In the *Build subgroups* widget we used the SD (Gamberger and Lavrač 2002) subgroup discovery algorithm with 20 % minimum support. In the *Select subgroups* widget we (arbitrarily) selected the subgroup $(Big\ spender = yes) \leftarrow (Cosenza = 0) \wedge (Gold = 1)$. This subgroup contains customers that are not from *Cosenza* and have a *Gold* bank account.

In the *Query data with subgroups, Table from sets of examples* and *Ranker* widgets we used the default settings. In the *SDM-Aleph* widget, we set the data format to 'list' and the cutoff parameter to 10 (this indicates that the input list will be split into two classes by the

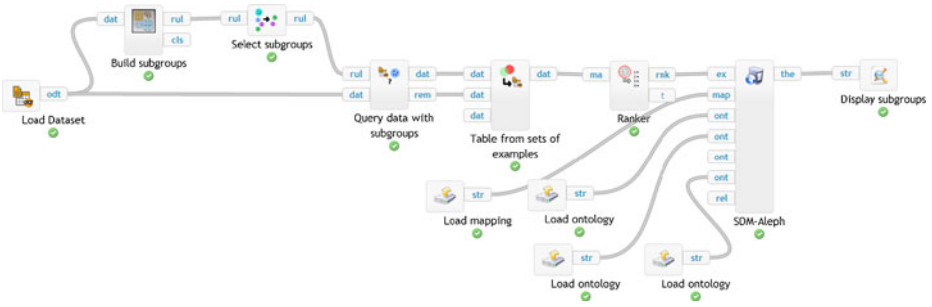

Fig. 3 The workflow implementing the solution to the motivational use case in ClowdFlows. The workflow can be found at http://clowdflows.org/workflow/1283/

Aleph system at the tenth attribute). This is necessary because the *Ranker* widget outputs a *list* of attributes and their scores; alternatively an Orange dataset can be used.

The best scoring subgroup found by *SDM-Aleph* is $top(X) \leftarrow Account(X)$. What is important to note from this simple example is that the *Account* ontological concept does not appear among the dataset attributes (leaves of the ontologies). This means that the explanation was found by generalizing the leaves (*Gold*, *Classic* and *Student* accounts) into the more general concept *Account* (see the first ontology in Fig. 2 to see the relation between these concepts). The explanation indicates that the main difference between the selected subgroup of customers and all other customers is in the type of the account they have. This is intuitive, since the selected subgroup contains the majority of customers with *Gold* accounts, while other customers have either *Classic* or *Student* accounts.

5.2 Biomedical use case

This subsection presents and discusses the application of the presented methodology on gene expression data. More specifically, we evaluate the methodology on the breast cancer dataset using our implementation of the methodology as a workflow in the ClowdFlows platform.

The gene expression dataset used in our analysis is the dataset published by Sotiriou et al. (2006) (GEO series GSE2990). It is a merge of the KJX64 and KJ125 datasets and contains expression values of 12,718 genes from 189 patients with primary operable invasive breast cancer. It also provides 22 metadata attributes such as age, grade, tumor size and survival time. We used the expert-curated re-normalized and binarized version of the dataset from the InSilico database (Taminau et al. 2011). Within the InSilico framework, the raw data was renormalized using fRMA (McCall et al. 2010) and a genetic barcode (0/1) was generated based on whether the expression of a gene was significantly higher (K standard deviations) than the no expression level estimated on a reference of approx. 800 samples. In this setting $g_i = 1$ means that gene g_i is over-expressed and $g_i = 0$ means that it is not. The ultimate goal of the experiment was to induce meaningful high-level semantic descriptions of subgroups found in the data which could provide important information in the clinical decision making process.

Our main motivation for developing the presented methodology is to descriptively characterize various breast cancer subtypes, while in the experiments presented here we focus on describing breast cancer grades, which enables us to focus on the evaluation of the methodology.

The conducted experiment on the presented dataset in the ClowdFlows environment employs processing components (widgets) in a complex data analysis workflow which is shown in Fig. 4.

In the first step, the *Load Dataset* widget is used to read the breast cancer patient data, i.e., a binarized version of the gene expression data (note that the frozen robust multiarray analysis (fRMA) normalization (McCall et al. 2010) is also available from the InSilico web page). As the GSE2990 dataset does not have pre-specified classes we have selected the *Grade* attribute as the target attribute using the *Select Attributes* widget. According to Elston and Ellis (1991) and Galea et al. (1992), histologic grade of breast carcinomas provides clinically important prognostic information. Approximately one half of all breast cancers are assigned histologic grade 1 or 3 status (low or high risk of recurrence) but a substantial percentage of tumors (30–60 %) are classified as histologic grade 2 (intermediate risk of recurrence) which is not informative for clinical decision making

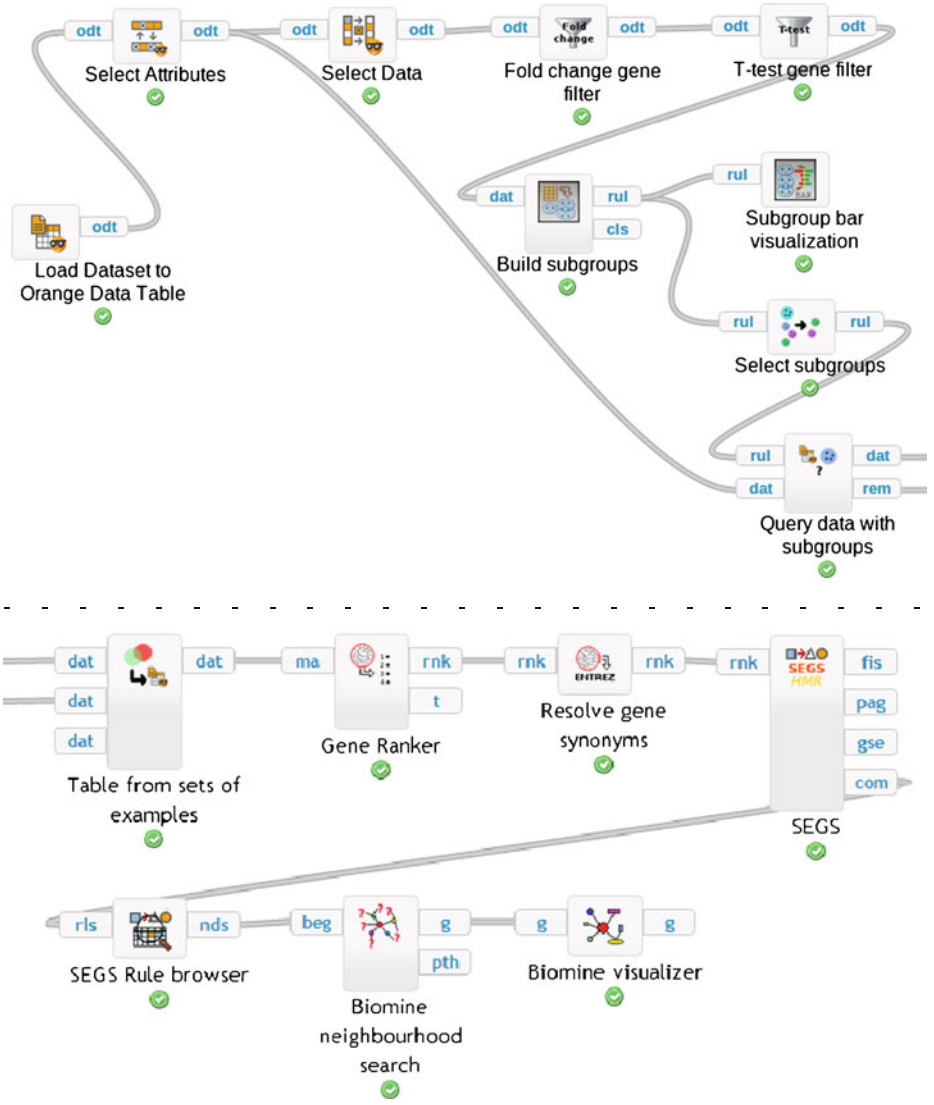

Fig. 4 A workflow implementing the proposed methodology in ClowdFlows (*first part*). The workflow was split into two parts in order to be more easily readable. The workflow can be found at http://clowdflows.org/workflow/911/

(Sotiriou et al. 2006). Obviously, to increase the prognostic value of tumor grading, further refinement of histologic grade 2 status is necessary (Sotiriou et al. 2006).

The third step of the workflow is to use the *Select data* widget to remove 17 unclassified examples for which the histologic grade is unknown. Although these examples may contain important information, this would require using unsupervised methods (e.g. clustering) instead of supervised subgroup discovery algorithms used in our experiments (note, however, that subgroup discovery in the presented workflow can easily be replaced by clustering or some other unsupervised method).

Next, attribute (gene) selection is performed using two gene filtering components which allow filtering the genes according to two scoring methods: fold change and t-test. Removal of unimportant genes is needed to reduce the search space of subgroup discovery methods to avoid the high-dimensionality problem. In our approach we have selected the genes in two stages: first, only the genes with a fold change of > 1 are selected, and second, only the genes with p-value $< 0, 01$ given by the t-test are selected. This yields a total of 399 genes to be used in the subgroup discovery process.

The *Build subgroups* widget implements SD (Gamberger and Lavrač 2002), APRIORI-SD (Kavšek and Lavrač 2006) and CN2-SD (Lavrač et al. 2004) subgroup discovery algorithms while the *Subgroup Bar visualization* component provides a facility of bar chart visualization, while the *Select subgroups* widget allows the selection of particular subgroups. The selected subgroups are used to query the original data (*Query data with subgroups*) to obtain the covered set of examples which are then merged with the rest of the data (*Table from sets of examples*). As a result it is possible to rank the genes in the re-constructed dataset according to their ability to differentiate between the discovered subgroups and the rest of the data. The ranking of genes is performed by the *Gene ranker* widget implementing the ReliefF algorithm.

Finally, the computed ranking is sent to the *SEGS* widget which calls the web service implementing the SEGS semantic subgroup discovery algorithm (SDM-SEGS and SDM-Aleph can also be used). As the SEGS algorithm has large time and space requirements it is implemented as a web service which allows it to run on a powerful server. SEGS induces rules providing explanations of the top ranked attributes by building conjunctions of ontology terms from the GO ontology, KEGG orthology, and interacting terms using the Entrez gene-gene interactions database as described in Section 3.3. In our experiments we have used the latest updates of the ontologies and annotations provided by NCBI[6] and the Gene Ontology project.

The subgroup discovery analysis yielded two large subgroups (Table 2) of Grade 3 patients. Using the GeneCards[7] on-line tool, we have confirmed that all of the genes from the subgroup descriptions are typically differentially expressed (up-regulated) in breast cancer tissue when compared with normal tissue.

In the rest of this section we focus on the larger subgroup #1, for which we have generated explanations (Table 3). A total of 90 explanations with p-value < 0.05 (estimated using permutation testing) were found. Due to space restrictions we display only the top 10 explanations generated by SEGS (for a complete list open the workflow from Fig. 4). For example, Explanation #1 describes genes which are annotated by GO/KEGG terms: *chromosome* and *cell cycle*.

In the study by Sotiriou et al. (2006) where the expression profiles of Grade 3 and Grade 1 patients were compared, the genes that are associated with histologic grade were shown to be mainly involved in cell cycle regulation and proliferation (uncontrollable division of cells is one of the hallmarks of cancer). The explanations of Subgroup #1 of Grade 3 patients in Table 3 agree with their findings. In general, the explanations describe genes that take part in cell cycle regulation (Explanations #1–#10), cell division (Explanation #3) and other components that indirectly affect cell division (e.g., Explanations #4 and #5: microtubules are structures that pull the cell apart when it divides).

[6] http://www.ncbi.nlm.nih.gov/gene

[7] http://www.genecards.org

Table 2 The best-scoring subgroups found using CN2-SD with default parameters for the Grade 3 patients

#	Subgroup description	TP	FP
1	Grade = 3 ← DDX39A = 1 ∧ DDX47 = 1 ∧ RACGAP1 = 1 ∧ ZWINT = 1 ∧ PITPNB = 1	43	5
2	Grade = 3 ← TPX2 = 1 ∧ DDX47 = 1 ∧ PITPNB = 1 ∧ HN1 = 1	26	0

TP and FP are the true positive and false positive rates, respectively

In our implementation, the user can choose one of the explanations (i.e., gene sets) to query the Biomine database (Eronen and Toivonen 2012). The Biomine engine offers advanced probabilistic graph searching techniques that can be used to find a neighborhood of the set of genes, or a graph connecting two separate gene sets. The result of both is a subgraph that can be explored with the *Biomine visualizer* widget. Figure 5 shows a part of the neighborhood graph for the gene set of explanation #2. In this particular case, the figure shows three types of nodes (*gene*, *biological process* and *pathway*) and the links between the nodes signify how are the nodes related (*participates in*, *codes for*).

To sum up, our study shows that by using our methodology one can automatically reproduce the observations noted in the earlier work by Sotiriou et al. This can encourage the researchers to apply the presented methodology in similar exploratory analytics tasks. Given the easy access and adaptability of the software the methodology can be simply reused in other domains, which is demonstrated in the next section on financial news articles.

5.3 Supporting factors comparison

In this subsection we present the results of the related *supporting factors* (Gamberger and Lavrač 2003) methodology, which is also used to characterize subgroups and can be of great help to domain experts in the interpretation of subgroup discovery patterns. Supporting factors are features that have statistically significantly different distributions in the positive examples of a selected subgroup, when compared to the control examples (negative cases) in the whole population and by themselves do not appear in the subgroup description. The difference is measured using the χ^2-test of independence.

Table 3 The explanations for the patients from subgroup #1 from Fig. 3

#	Explanation	*p*-value
1	chromosome ∧ cell cycle	0.000
2	cellular macromolecule metabolic process ∧ intracellular non-membrane-bounded organelle ∧ cell cycle	0.000
3	cell division ∧ nucleus ∧ cell cycle	0.000
4	regulation of mitotic cell cycle ∧ cytoskeletal part	0.000
5	regulation of mitotic cell cycle ∧ microtubule cytoskeleton	0.000
6	regulation of G2/M transition of mitotic cell cycle	0.000
7	regulation of cell cycle process ∧ chromosomal part	0.000
8	regulation of cell cycle process ∧ spindle	0.000
9	enzyme binding ∧ regulation of cell cycle process ∧ intracellular non-membrane-bounded organelle	0.000
10	ATP binding ∧ mitotic cell cycle ∧ nucleus	0.005

We omit the variables from the rules for better readability. Note that since the *p*-values are estimations, some can also have a value of 0

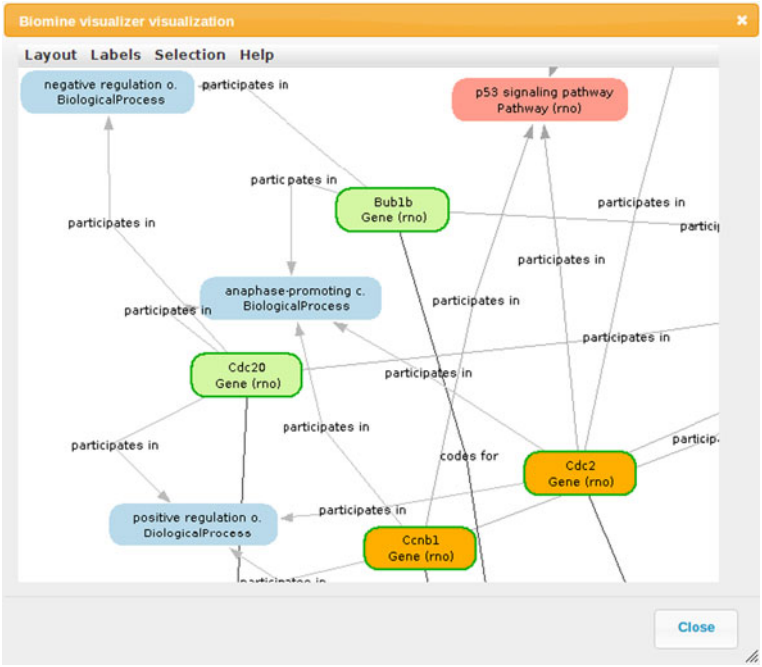

Fig. 5 The Biomine visualizer showing a part of a neighborhood graph for explanation #2 from Table 3

The methodology proposed in the present paper consists of several steps, where each step can be executed with multiple alternatives. The supporting factors methodology fits the best as a replacement to the last step. In this experiment, we assumed that all but the last step—explanation of subgroups—is the same as with our approach.

To be directly comparable to our results, we selected subgroup #1 (Fig. 2) as the target subgroup to characterize using supporting factors. We used a confidence level of 99 % ($p = 0.01$) and we report the best 10 supporting factors (Table 4).

The main diffrence that we can see is that supporting factors are represented as single genes, and the technique does not try to generalize over the genes and associate them with concepts from the gene ontology. This can of course be desirable for many use cases, but in this genomics experiment the characterization is not instantly obvious, since an additional look-up of individual genes is required by the domain expert.

On the other hand, the reported genes reaffirm the higher-level explanations produced by our methodology. For example, again using the GeneCards tool, we can find that the TPX2 gene is required for normal assembly of microtubules during apoptosis (cell death).

Table 4 Subgroup #1 from Fig. 2 and its top 10 supporting factors calculated with a confidence value of 99 % ($p = 0.01$)

Subgroup description	Supporting factors
Grade $= 3 \leftarrow$ DDX39A $= 1 \wedge$ DDX47 $= 1$ $\wedge$RACGAP1 $= 1 \wedge$ ZWINT $= 1 \wedge$ PITPNB $= 1$	TPX2, MAD2L1, CCNB2, CDK1, NUSAP1, CENPA, SNRPD1, GINS1, ASPM, PRC1

CCNB2 plays a key role in the control of the cell cycle and NUSAP1 is another microtubule-associated protein. The GINS1 plays an essential role in the initiation of DNA replication.

To sum up, the supporting factors approach can be important in domains where extra supportive evidence is needed (e.g., medical decision support), since it lists specific features that support a given subgroup. On the other hand, it does not provide a more general context, such as is possible using semantic subgroup discovery methods. Of course, the expert could also benefit from using these two methodologies side-by-side, since they characterize subgroups at two different levels of abstraction.

6 Conclusions

In this paper we presented a methodology for explaining subgroups or sets of instances using higher-level ontological concepts. First, a subgroup of instances is identified (e.g., using subgroup discovery or clustering), which is then characterized using ontological concepts thus providing insight into the main differences between the given subgroup and the remaining data.

We made the developed tools available for the ClowdFlows platform. Due to this implementation the tools are easily accessible, since ClowdFlows requires only an internet connection and a web browser.

As demonstrated by the two use cases, the proposed approach is general and can be employed in any application area, provided the existence of available domain ontologies and annotated data to be analyzed. In this paper, the real-life use case is from the genomics domain.

As the experts assume that there are several molecular subtypes of breast cancer, our main research interest of the genomics use case is to employ the presented methodology to descriptively characterize the hypothesized cancer subtypes. Hence, in addition to distinguishing between patients with breast cancer (the positive cases) and healthy patients, the challenge is to identify breast cancer subtypes by finding subgroups of patients which would be explained by the same gene functions, processes in which the genes interact. The approach presented in this paper has the potential of discovering groups of patients which correspond to the subtypes while explaining them using ontology terms describing gene functions, processes and pathways; in this paper, we applied the methodology to describe breast cancer grades with the aim of evaluation.

Using subgroup discovery we have identified two main subgroups that characterize Grade 3 breast cancer patients. These were then additionally explained using Gene Ontology concepts and KEGG pathways and the explanations (rules or subgroup descriptions of gene sets) agree with previous findings characterizing grades using microarray profiling.

Furthermore, compared to the related supporting factors approach, which is also used to characterize subgroups, the experiments show that supporting factors are more useful when concentrating on specific low-level attributes or features is desirable, but when more general descriptions are needed, they are not as easy to interpret as the method presented in this paper. This restriction is even more apparent on gene expression data, since the supporting factors are given in terms of single genes.

The results of the conducted experiments show the capabilities of the presented approach. In further work we will employ the methodology to detecting and characterizing subtypes of breast cancer. In further work, we will apply this methodology to other domains, as well as advance the level of exploitation of domain ontologies for providing explanations of the results of data mining.

Acknowledgments This work was supported by the Slovenian Ministry of Higher Education, Science and Technology [grant number P-103], the Slovenian Research Agency [grant number PR-04431], the SemDM project (Development and application of new semantic data mining methods in life sciences) [grant number J2-5478] and the FP7 European Commission project MUSE (Machine understanding for interactive storytelling) [grant number 296703].

References

Angiulli, F., Fassetti, F., Palopoli, L. (2013). Discovering characterizations of the behavior of anomalous subpopulations. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1280–1292. doi:10.1109/TKDE.2012.58.

Atzmüller, M., & Puppe, F. (2006). SD-Map—a fast algorithm for exhaustive subgroup discovery. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases (PKDD '06)* (pp. 6–17). Springer.

Bay, S.D., & Pazzani, M.J. (2001). Detecting group differences: mining contrast sets. *Data Mining and Knowledge Discovery*, *5*(3), 213–246.

Demšar, J., Zupan, B., Leban, G. (2004). *Orange: from experimental machine learning to interactive data mining, white paper*. Faculty of Computer and Information Science, University of Ljubljana. www.ailab.si/orange.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99)* (pp. 43–52).

Elston, C.W., & Ellis, I.O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, *19*(5), 403–410.

Eronen, L., & Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, *13*, 119.

Galea, M., Blamey, R., Elston, C., Ellis, I. (1992). The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, *22*, 207–219.

Gamberger, D., & Lavrač, N. (2002). Expert-guided subgroup discovery: methodology and application. *Journal of Artificial Intelligence Research (JAIR)*, *17*, 501–527.

Gamberger, D., & Lavrač, N. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, *28*(1), 27–57.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explor Newsl*, *11*, 10–18.

Hilario, M., Nguyen, P., Do, H., Woznica, A., Kalousis, A. (2011). Ontology-based meta-mining of knowledge discovery workflows. In N. Jankowski, W. Duch, K. Grabczewski (Eds.), *Meta-learning in computational intelligence, studies in computational intelligence* (Vol. 358, pp. 273–315). Berlin Heidelberg: Springer.

Jovanoski, V., & Lavrač, N. (2001). Classification rule learning with APRIORI-C. In P. Brazdil & A. Jorge (Eds.), *EPIA, lecture notes in computer science* (Vol. 2258, pp. 44–51). Berlin Heidelberg: Springer.

Kavšek, B., & Lavrač, N. (2006). APRIORI-SD: adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, *20*(7), 543–583.

Klösgen, W. (1996). Explora: a multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, (pp. 249–271). Menlo Park: American Association for Artificial Intelligence.

Kralj Novak, P., Lavrač, N., Webb, G.I. (2009). Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, *10*, 377–403.

Kranjc, J., Podpečan, V., Lavrač, N. (2012). Clowdflows: a cloud based scientific workflow platform. In P.A. Flach, T.D. Bie, N. Cristianini (Eds.), *ECML/PKDD (2), lecture notes in computer science* (Vol. 7524, pp. 816–819). Berlin Heidelberg: Springer.

Langohr, L., Podpečan, V., Petek, M., Mozetič, I., Gruden, K., Lavrač, N., Toivonen, H. (2013). Contrasting subgroup discovery. *Computer Journal*, *56*(3), 289–303.

Lavrač, N., Kavšek, B., Flach, P.A., Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, *5*, 153–188.

Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I., Kralj Novak, P. (2011). Using ontologies in semantic data mining with SEGS and g-SEGS. In *Proceedings of the international conference on discovery science* (*DS '11*) (pp. 165–178). Springer.

Lawrynowicz, A., & Potoniec, J. (2011). Fr-ont: an algorithm for frequent concept mining with formal ontologies. In M. Kryszkiewicz, H. Rybinski, A. Skowron, Z.W. Ras (Eds.), *ISMIS, lecture notes in computer science* (Vol. 6804, pp. 428–437). Berlin Heidelberg: Springer.

Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T. (2005). Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, *33*(Database issue).

McCall, M.N., Bolstad, B.M., Irizarry, R.A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics*, *11*(2), 242–253.

Podpečan, V., Juršič, M., Žakova, M., Lavrač, N. (2009). Towards a service-oriented knowledge discovery platform. In V. Podpečan & N. Lavrač (Eds.), *Third-generation data mining: towards service-oriented knowledge discovery* (pp. 25–36).

Podpečan, V., Lavrač, N., Mozetič, I., Kralj Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K. (2011a). SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, *12*, 416.

Podpečan, V., Zemenova, M., Lavrač, N. (2011b). Orange4WS environment for service-oriented data mining. *The Computer Journal*. doi:10.1093/comjnl/bxr077. Accessed 7 Aug 2011.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, *53*, 23–69.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, *98*(4), 262–272.

Srinivasan, A. (2007). *Aleph manual*. http://www.cs.ox.ac.uk/activities/machinelearning/Aleph/.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15,545–15,550.

Suzuki, E. (1997). Autonomous discovery of reliable exception rules. In *Proceedings of the third international conference on knowledge discovery and data mining* (pp. 259–262).

Suzuki, E. (2006). Data mining methods for discovering interesting exceptions from an unsupervised table. *Journal of Universal Computer Science*, *12*(6), 627–653.

Taminau, J., Steenhoff, D., Coletta, A., Meganck, S., Lazar, C., de Schaetzen, V., Duque, R., Molter, C., Bersini, H., Nowé, A., Weiss Solís, D.Y. (2011). InSilicoDB: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics*. doi:10.1093/bioinformatics/btr529.

Trajkovski, I., Lavrač, N., Tolar, J. (2008). SEGS: search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, *41*(4), 588–601.

Vavpetič, A., & Lavrač, N. (2013). Semantic subgroup discovery systems and workflows in the SDM-Toolkit. *Computer Journal*, *56*(3), 304–320.

Vavpetič, A., Podpečan, V., Meganck, S., Lavrač, N. (2012). Explaining subgroups through ontologies. In P. Anthony, M. Ishizuka, D. Lukose (Eds.), *Proceedings of PRICAI, lecture notes in computer science* (Vol. 7458, pp. 625–636). Berlin Heidelberg: Springer.

Vavpetič, A., Novak, P.K., Grčar, M., Mozetič, I., Lavrač, N. (2013). Semantic data mining of financial news articles. In *Proceedings of the international conference on discovery science (DS '13)*. Springer.

Webb, G.I., Butler, S.M., Newlands, D. (2003). On detecting differences between groups. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining (KDD-03)* (pp. 256–265).

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the first European conference on principles of data mining and knowledge discovery (PKDD '97)* (pp. 78–87). Springer.

Žáková, M., Železný, F., García-Sedano, J.A., Tissot, C.M., Lavrač, N., Kremen, P., Molina, J. (2006). Relational data mining applied to virtual engineering of product designs. In *Proceedings of the 16th international conference on inductive logic programming (ILP'06)* (pp. 439–453). Berlin/Heidelberg, Germany, Santiago de Compostela, Spain: Springer-Verlag.