Knowl Inf Syst (2011) 29:495–525
DOI 10.1007/s10115-010-0356-2

REGULAR PAPER

An overview on subgroup discovery: foundations and applications

Francisco Herrera · Cristóbal José Carmona · Pedro González · María José del Jesus

Received: 25 March 2010 / Revised: 24 August 2010 / Accepted: 23 October 2010 /
Published online: 11 November 2010

Abstract Subgroup discovery is a data mining technique which extracts interesting rules with respect to a target variable. An important characteristic of this task is the combination of predictive and descriptive induction. An overview related to the task of subgroup discovery is presented. This review focuses on the foundations, algorithms, and advanced studies together with the applications of subgroup discovery presented throughout the specialised bibliography.

Keywords Subgroup discovery · Knowledge discovery

1 Introduction

The aim of this paper is to present an overview of subgroup discovery by analysing the main properties, models, quality measures and real-world problems solved by subgroup discovery approaches.

Subgroup discovery [70,108] is a broadly applicable data mining technique aimed at discovering interesting relationships between different objects in a set with respect to a specific property which is of interest to the user the target variable. The patterns extracted are normally represented in the form of rules and called subgroups [101].

F. Herrera
Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
e-mail: herrera@decsai.ugr.es

C. J. Carmona (✉) · P. González · M. J. del Jesus
Department of Computer Science, University of Jaen, Jaén, Spain
e-mail: ccarmona@ujaen.es

P. González
e-mail: pglez@ujaen.es

M. J. del Jesus
e-mail: mjjesus@ujaen.es

Previous techniques have not been able to achieve this propose. For example, predictive techniques maximise accuracy in order to correctly classify new objects, and descriptive techniques simply search for relations between unlabelled objects. The need for obtaining simple models with a high level of interest led to statistical techniques which search for unusual relations [70].

In this way, subgroup discovery is somewhere halfway between supervised and unsupervised learning [78]. It can be considered that subgroup discovery lies between the extraction of association rules and the obtaining of classification rules.

The paper is organised as follows. Section 2 introduces subgroup discovery, its positioning in data mining and its main elements; Sect. 3 reviews the most important quality measures used in subgroup discovery; Sect. 4 presents a historical revision of the subgroup discovery algorithms; Sect. 5 describes different studies related to subgroup discovery; Sect. 6 shows different analyses of real-world problems using algorithms of subgroup discovery; finally, concluding remarks are presented in Sect. 7. In Appendix A, a description for each open source software tool with subgroup discovery algorithms is performed.

2 Subgroup discovery

In the following subsections, the formal definition of the subgroup discovery task, the relation with other data mining tasks and the main elements of a subgroup discovery algorithm are depicted.

2.1 Definition of subgroup discovery

The concept of subgroup discovery was initially introduced by Kloesgen [70] and Wrobel [108], and more formally defined by Siebes [101] but using the name Data Surveying for the discovery of interesting subgroups. It can be defined as [109]:

> In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

Subgroup discovery attempts to search relations between different properties or variables of a set with respect to a target variable. Due to the fact that subgroup discovery is focused in the extraction of relations with interesting characteristics, it is not necessary to obtain complete but partial relations. These relations are described in the form of individual rules.

Then, a rule (R), which consists of an induced subgroup description, can be formally defined as [43,84]:

$$R : Cond \rightarrow Target_{value}$$

where $Target_{value}$ is a value for the variable of interest (target variable) for the subgroup discovery task (which also appears as $Class$ in the literature), and $Cond$ is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the $Target_{value}$.

As an example, let D be a data set with three variables $Age = \{Less\,than\,25,\ 25\,to\,60,$ $More\,than\,60\}$, $Sex = \{M,\ F\}$ and $Country = \{Spain,\ USA,\ France,\ German\}$, and

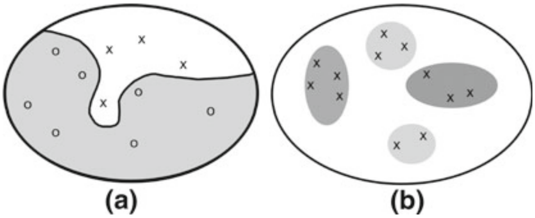

Fig. 1 Models of different techniques of knowledge discovery of databases

a variable of interest target variable $Money = \{Poor, \ Normal, \ Rich\}$. Some possible rules containing subgroup descriptions are:

$$R_1 : (Age = Less\,than\,25\,AND\,Country = German) \rightarrow \ Money = Rich$$
$$R_2 : (Age = More\,than\,60\,AND\,Sex = F) \rightarrow \ Money = Normal$$

where rule R_1 represents a subgroup of German people with less than 25 years old for which the probability of being rich is unusually high with respect to the rest of the population, and rule R_2 represents that women with more than 60 years old are more likely to have a normal economy than the rest of the population.

2.2 Subgroup discovery versus classification

Data mining is a stage of the Knowledge Discovery in Databases defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data" [41]. Description of the ten most used data mining algorithms can be found in [110]. Data mining techniques can be applied from two different perspectives:

- Predictive induction, whose objective is the discovery of knowledge for classification of prediction. Among its features, we can find classification [31], regression [31], or temporal series [21].
- Descriptive induction, whose main objective is the extraction of interesting knowledge from the data. Its features include association rules [2], summarisation [116] or subgroup discovery [70,108] can be mentioned.

In Fig. 1, the main difference between descriptive and predictive induction can be found. Figure 1a represents a precise and complex model (classifier) for predictive induction which divides perfectly the space in two determined regions with respect to the type of objects in the set. This model is based on precision and interpretability. However, Fig. 1b represents a model for descriptive induction which describes groups of elements (clusters) in the set, without a target variable, and based on support and confidence of the objects. As can be observed, the model on the left (predictive induction) has a different goal with respect to the model on the right (descriptive induction). Therefore, different heuristics and evaluation criteria in both types of learning are employed.

Subgroup discovery [70] is a technique for the extraction of patterns, with respect to a property of interest in the data, or target variable. This technique is somewhere halfway between predictive and descriptive induction, and its goal is to generate in a single and interpretable way subgroups to describe relations between independent variables and a certain value of the target variable. The algorithms for this task must generate subgroups for each value of the target variable. Therefore, an execution for each value of the variable must be performed.

A rule for subgroup discovery is represented in Fig. 2, where two values for the target variable can be found ($Target_{value} = x$ and $Target_{value} = o$). In this representation, a subgroup

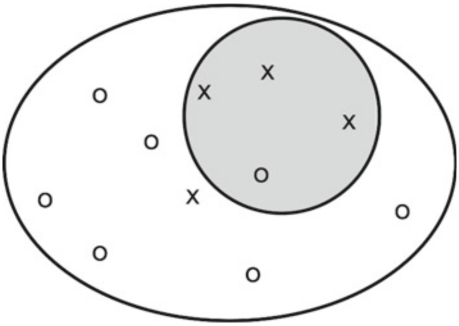

Fig. 2 Representation of a subgroup discovery rule with respect to a value (x) of the target variable

for the first value of the target variable can be observed, where the rule attempts to cover a high number of objects with a single function: a circle. As can be observed, the subgroup does not cover all the examples for the target value *x* even the examples covered are not positive in all the cases, but the form of this function is uniform and very interpretable with respect others. In this way, the algorithm achieves a reduction of the complexity. Furthermore, the true positive rate for the value of the target variable is high, with a value of 75%.

The subgroup discovery task is differentiated from classification techniques basically because subgroup discovery attempts to describe knowledge for the data while a classifier attempts to predict it. Furthermore, the model obtained by a subgroup discovery algorithm is usually simple and interpretable, while that obtained by a classifier is complex and precise.

Currently, several techniques lie halfway between descriptive and predictive data mining. “Supervised Descriptive Rule Induction” [78] is a new recently proposed paradigm which includes techniques combining the features of both types of induction, and its main objective is to extract descriptive knowledge from the data of a property of interest. These techniques use supervised learning to solve descriptive tasks. Within this new paradigm, the following data mining techniques are included:

– Subgroup Discovery [70,108], defined as the extraction of interesting subgroups for a target value.
– Contrast Set Mining [17], defined as “a conjunction of attribute-value pairs defined on groups with no attribute occurring more than once”.
– Emerging Pattern Mining [38] defined as “patterns whose frequencies in two classes differ by a large ratio”.

The main difference between these techniques is that while subgroup discovery algorithms attempt to describe unusual distributions in the search space with respect a value of the target variable, contrast set and emerging pattern algorithms seek relationships of the data with respect to the possible values of the target variable. Contrast set algorithms attempts to obtain high differences of support between the possible values and emerging pattern algorithms search patterns with different frequencies in two classes of the target variable. These last two techniques are based on measures of coverage and accuracy and subgroup discovery is also focused on novelty and unusualness measures as can be observed in the following sections.

2.3 Main elements in a subgroup discovery algorithm

Different elements can be considered the most important when a subgroup discovery approach must be applied. These elements are defined below [13]:

- *Type of the target variable* Different types for the variable can be found: binary, nominal or numeric. For each one, different analyses can be applied considering the target variable as a dimension of the reality to study.
    - Binary analysis. The variables have only two values (True or False), and the task is focused on providing interesting subgroups for each of the possible values.
    - Nominal analysis. The target variable can take an undetermined number of values, but the philosophy for the analysis is similar to the binary, to find subgroups for each value.
    - Numeric analysis. This type is the most complex because the variable can be studied different ways such as dividing the variable in two ranges with respect to the average, discretisising the target variable in a determined number of intervals [91], or searching for significant deviations of the mean among others.
- *Description language* The representation of the subgroups must be suitable for obtaining interesting rules. These rules must be simple and therefore are represented as attribute-value pairs in conjunctive or disjunctive normal form in general. Furthermore, the values of the variables can be represented as positive and/or negative, through fuzzy logic, or through the use of inequality or equality and so on.
- *Quality measures* These are a key factor for the extraction of knowledge because the interest obtained depends directly on them. Furthermore, quality measures provide the expert with the importance and interest of the subgroups obtained. Different quality measures have been presented in the specialised bibliography [45,70,74,84], but there is no consensus about which are the most suitable for use in subgroup discovery. In Sect. 3, the most important quality measures for subgroup discovery are presented.
- *Search strategy* This is very important, since the dimension of the search space has an exponential relation to the number of features and values considered. Different strategies have been used up to the moment, for example beam search, evolutionary algorithms, search in multi-relational spaces, etc. The algorithms implemented and their search strategies are shown in Sect. 4.

3 Quality measures of subgroup discovery

One of the most important aspects in subgroup discovery is the choice of the quality measures employed to extract and evaluate the rules. There is no current consensus in the field about which are the most suitable for both processes, and there are a wide number of measures presented throughout the bibliography.

The most common quality measures used in subgroup discovery are described here, classified by their main objective such as complexity, generality, precision, and interest.

3.1 Measures of complexity

These measures are related to the interpretability of the subgroups, i.e. to the simplicity of the knowledge extracted from the subgroups. Among them can be found:

- *Number of rules (n_r)*: It measures the number of induced rules.
- *Number of variables (n_v)*: It measures the number of variables of the antecedent. The number of variables for a set of rules is computed as the average of the variables for each rule of that set.

3.2 Measures of generality

These quality measures are used to quantify the quality of individual rules according to the individual patterns of interest covered. The quality measures of this type can be observed below:

- *Coverage*: It measures the percentage of examples covered on average [85]. This can be computed as:

$$Cov(R) = \frac{n(Cond)}{n_s} \quad (1)$$

  where n_s is the number of total examples and $n(Cond)$ is the number of examples which satisfy the conditions determined by the antecedent part of the rule.
- *Support*: It measures the frequency of correctly classified examples covered by the rule [85]. This can be computed as:

$$Sup(R) = \frac{n(Target_{value} \cdot Cond)}{n_s} \quad (2)$$

  where $n(Target_{value} \cdot Cond) = TP$ is the number of examples which satisfy the conditions and also belong to the value for the target variable in the rule.

3.3 Measures of precision

These quality measures show the precision of the subgroups and are widely used in the extraction of association rules and classification. Within this group can be found:

- *Confidence*: It measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. This can be computed with different expressions, e.g. [3]:

$$Cnf(R) = \frac{n(Target_{value} \cdot Cond)}{n(Cond)} \quad (3)$$

  This quality measure can also be found as *accuracy* in the specialised bibliography. In [61], an expression adapted for fuzzy rules can be found.
- *Precision measure Q_c*: It measures the tradeoff between the true and false positives covered in a lineal function [43]. This can be computed as:

$$Q_c(R) = TP - (c \cdot FP) = n(Target_{value} \cdot Cond) - c \cdot n(\overline{Target_{value}} \cdot Cond) \quad (4)$$

  where $n(\overline{Target_{value}} \cdot Cond) = FP$ are the examples satisfying the antecedent but not the target variable, and the parameter c is used as a generalisation parameter. This quality measure is easy to use because of the intuitive interpretation of this parameter.
- *Precision measure Q_g*: It measures the tradeoff of a subgroup between the number of examples classified perfectly and the unusualness of their distribution [70]. This can be computed as:

$$Q_g(R) = \frac{TP}{FP + g} = \frac{n(Target_{value} \cdot Cond)}{n(\overline{Target_{value}} \cdot Cond) + g} \quad (5)$$

  where g is used as a generalisation parameter, usually configured between 0.50 and 100. This concept was the first used for measuring the quality of the subgroups.

A modified Q_g *with weighting* can be observed in [43]. This measure introduces the concept of weighting with respect to the examples of the database and can be computed as:

$$Q'_g(R) = \frac{\sum_{TP} \frac{1}{c(e)}}{FP + g} = \frac{\sum_{n(Target_{value} \cdot Cond)} \frac{1}{c(e)}}{n(\overline{Target_{value}} \cdot Cond) + g} \tag{6}$$

where $c(e)$ measures how many times the example has been covered by any rule. This quality measure is used in an iterative process for obtaining subgroups. Initially, the value of $c(e)$ is 1, and when the example is covered by any rule this value is incremented by 1. In this way, the rules which cover new examples do not lose value.

3.4 Measures of interest

These measures are intended for selecting and ranking patterns according to their potential interest to the user. Within this type of measures can be found:

- *Interest*: It measures the interest of a rule determined by the antecedent and consequent [93]. It can be computed as:

$$Int(R) = \frac{\sum_{i=1}^{n_v} Gain(A_i)}{n_v \cdot \log_2 (|dom(G_k)|)} \tag{7}$$

  where $Gain$ is the information gain, $|dom(G_k)|$ is the cardinality of the target variable, and A_i is the number of values or intervals of the variable.
- *Novelty*: This measure is able to detect unusual subgroups [108]. It can be computed as:

$$Nov(R) = n(Target_{value} \cdot Cond) - (n(Target_{value}) \cdot n(Cond)) \tag{8}$$

  where $n(Target_{value})$ are all the examples of the target variable.
- *Significance*: This measure indicates the significance of a finding, if measured by the likelihood ratio of a rule [70].

$$Sig(R) = 2 \cdot \sum_{k=1}^{n_c} n(Target_{valuek} \cdot Cond) \cdot log \frac{n(Target_{valuek} \cdot Cond)}{n(Target_{valuek}) \cdot p(Cond)} \tag{9}$$

  where $p(Cond)$, computed as $n(Cond)/n_s$, is used as a normalised factor, and n_c is the number of values of the target variable. It must be noted that although each rule is for a specific $Target_{value}$, the significance measures the novelty in the distribution impartially, for all the values.

3.5 Hybrid measures

In this group can be found a large number of quality measures because subgroup discovery attempts to obtain a tradeoff between generality, interest and precision in the results obtained. The different quality measures used can be found below:

- *Sensitivity*: This measure is the proportion of actual matches that have been classified correctly [70]. It can be computed as:

$$Sens(R) = TPr = \frac{TP}{Pos} = \frac{n(Target_{value} \cdot Cond)}{n(Target_{value})} \tag{10}$$

where Pos are all the examples of the target variable ($n(Target_{value})$). This quality measure was used in [61] as *Support based on the examples of the class* and used to evaluate the quality of the subgroups in the Receiver Operating Characteristic (ROC) space. Sensitivity combines precision and generality related to the target variable.

- *False Alarm*: This measure is also known as the false-positive rate [43]. It covers to the examples which are not in the target variable and can be computed as:

$$FA(R) = FPr = \frac{FP}{Neg} = \frac{n(\overline{Target_{value}} \cdot Cond)}{n(\overline{Target_{value}})} \tag{11}$$

  where Neg is number of the examples that are not part of the $Target_{value}$, i.e. $n(\overline{Target_{value}})$. This quality measure is used for evaluating the quality of the subgroups in the ROC space.

- *Specificity*: It measures the proportion of negative cases incorrectly classified [70]. It can be computed as:

$$Spec(R) = \frac{TN}{TN + FP} = \frac{TN}{Neg} = \frac{n(\overline{Target_{value} \cdot Cond})}{n(\overline{Target_{value}})} \tag{12}$$

  where $n(\overline{Target_{value} \cdot Cond})$ are the examples which do not satisfy both condition and consequent.

- *Unusualness*: This measure is defined as the weighted relative accuracy of a rule [81]. It can be computed as:

$$WRAcc(R) = \frac{n(Cond)}{n_s}\left(\frac{n(Target_{value} \cdot Cond)}{n(Cond)} - \frac{n(Target_{value})}{n_s}\right) \tag{13}$$

  The unusualness of a rule can be described as the balance between the coverage of the rule $p(Cond_i)$ and its accuracy gain $p(Target_{value} \cdot Cond) - p(Target_{value})$. This quality measure is derived from *novelty* (8).

  A modified unusualness based on the weights of the examples is presented in [82] and can be computed as:

$$WRAcc'(R) = \frac{n'(Cond)}{n'_s}\left(\frac{n'(Target_{value} \cdot Cond)}{n'(Cond)} - \frac{n'(Target_{value})}{n'_s}\right) \tag{14}$$

  where n'_s is the sum of the weights of all examples, $n'(Cond)$ is the sum of the weights of all covered examples, and $n'(Target_{value} \cdot Cond)$ is the sum of the weights of all correctly covered examples.

  Several $WRAcc$ variants have been presented in [1] for measuring problems with multiple values in the $Target_{variable}$.

In Table 1, a summary of the quality measures shown previously can be found. In this table, the name of the $Quality\ measure$ and its main characteristics (highlighted with X) are summarised.

4 Historical revision of models in subgroup discovery

In this section, the subgroup discovery approaches developed so far are described. There are several proposals of algorithms for subgroup discovery. To classify these algorithms, it can be distinguished between extensions of classification algorithms, extensions of association

Table 1 Classification of the most important quality measures used in subgroup discovery

Quality measure		Complexity	Generality	Precision	Interest
n_r	Number of rules	X			
n_v	Number of variables	X			
Cov (1)	Coverage		X		
Sup (2)	Support		X		
Cnf (3)	Confidence			X	
Q_c (4)	Precision measure Q_c			X	
Q_g (5)	Precision measure Q_g			X	
Int (7)	Interest				X
Nov (8)	Novelty				X
Sig (9)	Significance				X
$Sens$ (10)	Sensitivity		X	X	
FA (11)	False Alarm		X	X	
$Spec$ (12)	Specificity		X	X	
$WRAcc$ (13)	Unusualness		X	X	X

Table 2 Classification of the algorithms for subgroup discovery developed so far

Extensions of classification algorithms
EXPLORA [70]
MIDOS [108]
SubgroupMiner [72]
SD [43]
CN2-SD [85]
RSD [83,112]
Extensions of association algorithms
APRIORI-SD [66,68]
SD4TS [92]
SD-MAP [10]
DpSubgroup [55]
Merge-SD [54]
IMR [20]
Evolutionary algorithms
SDIGA [61]
MESDIF [18,60]
NMEEF-SD [27,28]

algorithms and evolutionary fuzzy systems. Table 2 shows the main algorithms for subgroup discovery developed so far under this classification.

In addition to the algorithms given in Table 2, there are additional approaches that involve modifications or extensions of these algorithms. All these algorithms are detailed in the following subsections, comparing their characteristics.

Table 3 Features of the pioneering algorithms for subgroup discovery

Algorithm	Type of target variable	Description language	Quality measures	Search strategy
EXPLORA [70]	Categorical	Conjunctions of pairs attribute-value. Operators = and ≠	Evidence, generality, redundancy or simplicity among others [70]	Exhaustive and heuristic without pruning
MIDOS [108]	Binary	Conjunctions of pairs. Operators $=, <, >$ and ≠	Novelty (8) or distributional unusualness [109] among others	Exhaustive and minimum support pruning

4.1 Extensions of classification algorithms

Among the algorithms for subgroup discovery developed as extensions of classification algorithms, it can be distinguished between the pioneering algorithms and those based on classification algorithms. Both are described below.

4.1.1 The pioneering algorithms

The first algorithms developed for subgroup discovery—EXPLORA and MIDOS—are extensions of classification algorithms and use decision trees. They can employ two different strategies for the search, exhaustive and heuristic search, and several quality measures to evaluate the quality of the subgroups.

- EXPLORA [70] was the first approach developed for subgroup discovery. It uses decision trees for the extraction of rules. The rules are specified by defining a descriptive schema and implementing a statistical verification method. The interest of the rules is measured using statistical measures [70] such as evidence, generality, redundancy or simplicity, among others. EXPLORA can apply both exhaustive and heuristic subgroup discovery strategies without pruning.
- MIDOS [108] applies the EXPLORA approach to multi-relational databases. The goal is to discover subgroups of the target variable (defined as first-order conjunctions) that have an unusual statistical distribution with respect to the complete population. MIDOS uses optimistic estimation and searches the space of rules exhaustively, except for safe pruning (using a minimum support pruning) for binary target variables. In order to find precise and significant subgroups, the size of the subgroups and the distributional unusualness are considered. This algorithm can also use sampling in the example space to reduce the search space and speed up the search process.

These algorithms can use exhaustive or heuristic search. Exhaustive evaluation of the candidate rules allows the best subgroups to be found, but if the search space becomes too large this is not affordable. Then a heuristic search can be used to reduce the number of potential subgroups to consider (if the rules are ordered by generality, parts of the search space can be pruned). EXPLORA performs the subgroup discovery task from data in a single relation, while MIDOS can handle multiple relational tables. Table 3 summarises the features of both algorithms.

4.1.2 Algorithms based on classification

Several algorithms developed for the subgroup discovery task have been developed by means of adaptations of classification rule learners. Classification rule learning algorithms have the objective of generating models consisting of a set of rules inducing properties of all the classes of the target variable, while in subgroup discovery the objective is to discover individual rules of interest. Moreover, classification rule learning algorithms do not appropriately address the subgroup discovery task as they use the covering algorithm for rule set construction. So, in order to use a classification rule learning algorithm for subgroup discovery, some modifications must be implemented. The algorithms described here attempt to overcome the inappropriate bias of the standard covering algorithm (only the first induced rules may be of interest as subgroup descriptions). They then use a modified weighted covering algorithm and introduce example weights to modify the search heuristic. They are briefly detailed below:

- SubgroupMiner [72] is an extension of EXPLORA and MIDOS. It is an advanced subgroup discovery system that uses decision rules and interactive search in the space of the solutions, allowing the use of very large databases by means of the efficient integration of databases, multi-relational hypotheses, visualisation based on interaction options, and the discovery of structures of causal subgroups. This algorithm can use several quality measures to verify if the statistical distribution of the target is significantly different in the extracted subgroup, but the most usual is the binomial test [70]. This can handle both numeric and nominal target attributes, but for numeric variables a previous discretisation is performed.
- SD [43] is a rule induction system based on a variation of the beam search algorithms and guided by expert knowledge: instead of defining an optimal measure to discover and automatically select the subgroups, the objective is to help the expert in performing flexible and effective searches on a wide range of optimal solutions. Discovered subgroups must satisfy the minimal support criteria and must also be relevant. The algorithm keeps the best subgroups descriptions in a fixed width beam and in each iteration a conjunction is added to every subgroup description in the beam, replacing the worst subgroup in the beam by the new subgroup if it is better. To evaluate the quality of the subgroups, Q_g measure (5) is used. In [44], other quality measures based on the results of the contingency table are presented: sensitivity (10), specificity (12), false alarm (11), support (2), and confidence (3) to measure the subgroups with more quality and accuracy. High quality subgroups must cover as many $Target_{value}$ examples and as few non-$Target_{value}$ examples as possible. The algorithm SD uses a visualisation method [48] to provide the experts with an easy tool to test the subgroups. In [45], methods to avoid noise and outliers values are presented. Different constraints (filtering of subgroups) are described for the algorithm SD in [80]: Maximum length (a threshold defined by the user) and minimum support for the rules. This algorithm is implemented in a module of the software tool Orange, which is briefly described in Appendix A.1.
- CN2-SD [84] is a subgroup discovery algorithm obtained by adapting a standard classification rule learning approach CN2 [32,33] to subgroup discovery. It induces subgroups in the form of rules using a modified unusualness (13) as the quality measure for rule selection. This approach performs subgroup discovery through the following modifications of CN2: (a) replacing the accuracy-based search heuristic with a new unusualness heuristic that combines the generality and accuracy of the rule; (b) incorporating example weights into the covering algorithm; (c) incorporating example weights into the

Table 4 Features of the subgroup discovery algorithms based on classification rule learners

Algorithm	Type of target variable	Description language	Quality measures	Search strategy
SubgroupMiner [72]	Categorical	Conjunctions of pairs attribute-value. Operator $=$	Binomial test [70]	Beam search
SD [43]	Categorical	Conjunctions of pairs. Operators $=$, $<$ and $>$	Q_g (5)	Beam search
CN2-SD [84]	Categorical	Conjunctions of pairs. Operators $=$, $<$, $>$ and $\neq$	Unusualness (13)	Beam search
RSD [83,112]	Categorical	Conjunctions of first-order features. Operators $=$, $<$ and $>$	Unusualness (13), significance (9) or coverage (1)	Beam search

unusualness search heuristic; (d) using probabilistic classification based on the class distribution of examples covered by individual rules. An extension of this algorithm named CN2-MSD [1] has been developed to manage multi-class target variables. This algorithm is implemented in a module of the software tool Orange (see Appendix A.1).

- RSD (Relational subgroup discovery) [83,112] has the objective of obtaining population subgroups which are as large as possible, with a statistical distribution as unusual as possible with respect to the property of interest, and different enough to cover most of the target population. It is an upgrade of the CN2-SD algorithm which enables relational subgroup discovery.

SubgroupMiner is the first algorithm which considers the use of numerical target variables, though it is necessary to perform a previous discretisation. The algorithms SD, CN2-SD and RSD use different heuristics for the subgroup discovery task. By definition, the interestingness of a subgroup depends on its unusualness and size; therefore, the rule quality evaluation heuristics need to combine both factors. CN2-SD and RSD use the unusualness, and SD uses the generalisation quotient. Moreover, SD and CN2-SD are propositional, while RSD is a relational subgroup discovery algorithm. Table 4 summarises the features of algorithms based on classification rule learners.

4.2 Extensions of association algorithms

An association rule algorithm attempts to obtain relations between the variables of the data set. In this case, several variables can appear both in the antecedent and consequent of the rule. In contrast, in subgroup discovery the consequent of the rule, consisting of the property of interest is prefixed. The characteristics of the association rule algorithms make it feasible to adapt these algorithms for the subgroup discovery task. The algorithms based on association rule learners are briefly described below:

- APRIORI-SD [66,68] is developed by adapting to subgroup discovery the classification rule learning algorithm APRIORI-C [63], a modification of the original APRIORI association rule learning algorithm [2]. APRIORI-SD uses a postprocessing mechanism, unusualness (13), as the quality measure for the induced rules and probabilistic classification of the examples. For the evaluation of the set of rules, the area under the

ROC curve is used, in conjunction with the support (2) and significance (9) of each individual rule, and the size and accuracy of the set of rules. This algorithm is implemented in a module of the software tool Orange, which is briefly described in Appendix A.1.

- SD4TS [92] is an algorithm based on APRIORI-SD but using the quality of the subgroup to prune the search space even more. The quality measure used is specified by the problem analysed. The algorithm does not need a covering heuristic.
- SD-Map [10] is an exhaustive subgroup discovery algorithm that uses the well-known FP-growth method [56] for mining association rules with adaptations for the subgroup discovery task. It uses an implicit depth-first search step for evaluating the subgroup hypotheses, included in the divide-and-conquer frequent pattern growth (that is, by the reordering/sorting optimisation). SD-Map uses a modified FP-growth step that can compute the subgroup quality directly without referring to other intermediate results. It can use several quality functions like Piatetsky-Shaphiro [70], unusualness (13), or the binomial test [70], among others. The adaptations of the algorithms based on APRIORI for subgroup discovery are also valid for the FP-growth method. This algorithm is implemented in the software tool VIKAMINE, which is briefly described in Appendix A.2. An extension named SD-Map$\star$ [8] has been developed which is applicable for binary, categorical, and continuous target variables.
- DpSubgroup [55] is a subgroup discovery algorithm that uses a frequent pattern tree to obtain the subgroups efficiently. It incorporates tight optimistic estimate pruning and focuses on binary and categorical target variables. DpSubgroup uses an explicit depth-first search step for evaluating the subgroup hypotheses. It makes use of the FpTree-based data representations introduced by SD-Map algorithm and focuses on binary and categorical target concepts.
- Merge-SD [54] is a subgroup discovery algorithm that prunes large parts of the search space by exploiting bounds between related numerical subgroup descriptions. In this way, the algorithm can manage data sets with numeric attributes. The algorithm uses a new pruning scheme which exploits the constraints among the quality of subgroups ranging over overlapping intervals. Merge-SD performs a depth-first search in the space of subgroup descriptions.
- IMR [20] is an alternative algorithmic approach for the discovery of non-redundant subgroups based on a breadth-first strategy. It searches for equivalence classes of descriptions with respect to their extension in the database rather than individual descriptions. So the algorithm searches in the space of subgroup extensions and has a potentially reduced search space returning at most one description of each extension to the user. It can use several quality measures, but in the experimentations, the one used is the binomial test. To manage continuous attributes, they must be previously discretised (the authors use the minimal entropy discretisation).

Some of these algorithms like APRIORI-SD or SD4TS are obtained from the adaptation to subgroup discovery of the association rule learner algorithm APRIORI, but others like SD-MAP, DpSubgroup or Merge-SD are adaptations of FP-Growth. All of them use decision trees for representation. Only Merge-SD and SD-MAP$\star$ can handle numeric or continuous variables, and for the rest a previous discretisation is necessary. Note that the discretisation scheme used affects the results obtained depending on the problem to be solved. Table 5 summarises the features of these algorithms.

Table 5 Features of the subgroup discovery algorithms based on association rule learners

Algorithm	Type of target variable	Description language	Quality measures	Search strategy
APRIORI-SD [66,68]	Categorical	Conjunctions of pairs attribute-value. Operators $=$, $<$ and $>$	Unusualness (13)	Beam search with minimum support pruning
SD4TS [92]	Categorical	Conjunction of pairs. Operators $=, <$ and $>$	Prediction quality [92]	Beam search with pruning
SD-MAP [10]	Binary	Conjunctive languages with internal disjunctions. Operator $=$	Piatetsky–Shapiro [70], unusualness (13), binomial test [70], among others	Exhaustive search with minimum support pruning
SD-MAP⋆ [8]	Continuous	Conjunctive languages with internal disjunctions. Operator $=$	Piatetsky–Shapiro [70], unusualness (13), lift [22]	Exhaustive search with minimum support pruning
DpSubgroup [55]	Binary and categorical	Conjunctions of pairs. Operator $=$	Piatetsky–Shaphiro [70], split, gini and pearson's χ^2 among others [55]	Exhaustive search with tight optimistic estimate pruning
MergeSD [54]	Continuous	Conjunctions of pairs. Operators $=, <, >, \neq$ and $intervals$	Piatetsky–Shapiro [70] among others	Exhaustive search with pruning based on constraints among the quality of subgroups
IMR [20]	Categorical	Conjunctions of pairs. Operator $=$	Binomial test [70]	Heuristic search with optimistic estimate pruning

4.3 Evolutionary algorithms for extracting subgroups

Subgroup discovery is a task which can be approached and solved as optimisation and search problems. Evolutionary algorithms [102] imitate the principles of natural evolution in order to form searching processes. One of the most widely used types of evolutionary algorithms are genetic algorithms, inspired by natural evolution processes and initially defined by Holland [53,58]. The heuristic used by this type of algorithm is defined by a fitness function, which determines which individuals (rules in this case) will be selected to form part of the new population in the competition process. This makes genetic algorithms very useful for the subgroup discovery task. The evolutionary algorithms proposed for extracting subgroups are explained below:

- SDIGA [61] is an evolutionary fuzzy rule induction system. It uses as quality measures for the subgroup discovery task adaptations of the measurements used in the association rules induction algorithms, confidence (3), and support (2) and can also use other measures such as interest (7), significance (9), sensitivity (10) or unusualness (13). The algorithm evaluates the quality of the rules by means of a weighted average of the measures selected. An analysis of different combinations of quality measures can be observed in [26]. SDIGA uses linguistic rules [59] as description language to specify the subgroups.

This algorithm is implemented in the software tool KEEL, which is briefly described in Appendix A.3.

- MESDIF [18,60] is a multi-objective genetic algorithm for the extraction of fuzzy rules which describe subgroups. The algorithm extracts a variable number of different rules expressing information on a single value of the target variable. The search is based on the multi-objective SPEA2 [118] approach, and so applies the concepts of elitism in the rule selection (using a secondary or elite population) and the search for optimal solutions in the Pareto front. It can use several quality measures at a time to evaluate the rules obtained, like confidence (3), support (2), (10), significance (9) or unusualness (13). This algorithm is implemented in the software tool KEEL (see Appendix A.3).
- NMEEF-SD [27,28] is an evolutionary fuzzy system whose objective is to extract descriptive fuzzy and/or crisp rules for the subgroup discovery task, depending on the type of variables present in the problem. NMEEF-SD has a multi-objective approach whose search strategy is based on NSGA-II [35], which is based on a non-dominated sorting approach, and on the use of elitism. This algorithm uses specific operators to promote the extraction of simple, interpretable, and high quality rules. It allows a number of quality measures to be used both for the selection and the evaluation of rules within the evolutionary process, including confidence (3), support (2), sensitivity (10), significance (9), and unusualness (13). This algorithm is implemented in the software tool KEEL (see Appendix A.3).

The evolutionary algorithms proposed so far for the subgroup discovery task are based on the hybridisation between fuzzy logic and evolutionary algorithms, known as evolutionary fuzzy systems [34,57]. They provide novel and useful tools for pattern analysis and for extracting new kinds of useful information.

As claimed in [39], “the use of fuzzy sets to describe associations between data extends the types of relationships that may be represented, facilitates the interpretation of rules in linguistic terms, and avoids unnatural boundaries in the partitioning of the attribute domains”. It is especially useful in domains where the boundaries of a piece of information used may not be clearly defined. The evolutionary algorithms allow the inclusion of quality measures in the process in order to obtain rules with suitable values not only in the selected quality measures but also in the others. The best way to obtain solutions with a good compromise between the quality measures for subgroup discovery is through a multi-objective evolutionary algorithm approach. Table 6 summarises the features of these proposals.

5 Related work in subgroup discovery

When applying subgroup discovery approaches, several aspects must be taken into account. In this section, we focus on describing the proposals related to the preprocessing of the data and postprocessing the knowledge, the discretisation of continuous variables, the use of domain knowledge, and the visualisation of the results.

5.1 Scalability in subgroup discovery

When applying data mining techniques to real-world problems, these usually have high dimensionality, unavoidable for most of the usual algorithms. There are two typical possibilities when a data mining algorithm does not work properly with high dimensional data sets

Table 6 Features of the evolutionary algorithms for extracting subgroups

Algorithm	Type of target variable	Description language	Quality measures	Search strategy
SDIGA [61]	Nominal	Conjunctive or disjunctive fuzzy rules. Operator =	Confidence (3) support (2), sensitivity (10), interest (7), significance (9) or unusualness (13) among others	Genetic algorithm
MESDIF [18,60]	Nominal	Conjunctive or disjunctive fuzzy rules. Operator =	Confidence (3) support (2), sensitivity (10), significance (9) or unusualness (13) among others	Multi-objective genetic algorithm
NMEEF-SD [27,28]	Nominal	Conjunctive or disjunctive fuzzy rules. Operator =	Confidence (3) support (2), sensitivity (10), significance (9) or unusualness (13) among others	Multi-objective genetic algorithm

[37]: redesigning the algorithm to run efficiently with huge input data sets or reducing the size of the data without changing the result drastically.

Sampling is one of the techniques most widely used in data mining to reduce the dimensionality of a data set and consists of the selection of particular instances of the data set according to some criterion. The application of a sampling technique in the initial database without considering dependencies and relationships between variables could lead to an important loss of knowledge for the subgroup discovery task. If it is necessary to apply some technique to scaling down the data set in a subgroup discovery algorithm, it is especially important to ensure that no important information for the extraction of interesting subgroups in the data is lost.

The following works related to scalability in subgroup discovery has been developed:

- In [98], a sampling algorithm, the Generic Sequential Sampling algorithm, is used to find the best hypothesis from a given space. The algorithm is sequential because it returns hypotheses which seem particularly good for the database.
- In [100], a method to filter the initial database in a distribution proportional to the initial was presented, allowing an improvement of the results obtained by the subgroup discovery approach (in this case the CN2-SD algorithm). The idea of the proposed method is "to construct samples that do not show the biases underlying previously discovered patterns".
- In [25], the authors tested whether instance selection alters the values of the quality measures of the subgroups. Later, the algorithm CN2-SD in combination with instance selection methods was run with large databases, concluding that the preprocessing methods can be applied before CN2-SD, because the results are maintained.
- In [24], a study was performed showing the benefits of using a combination of stratification and instance selection for scaling down the data set before applying the subgroup discovery algorithm APRIORI-SD to large databases. In this paper, two novel approaches for stratified instance selection based on increasing the presence of minority classes were presented.

5.2 Preprocessing of the variables

It is very common that some of the variables collected in the data sets used to apply subgroup discovery techniques are continuous variables. Most of the subgroup discovery algorithms are not able to handle continuous variables. In this case, a previous discretisation can be applied using different mechanisms [40,88]. Different preprocessing techniques in subgroup discovery can be observed:

- Some approaches can manage continuous variables in the condition of the rule (explaining variables) without the need of a previous discretisation. This is performed in the evolutionary algorithms proposed in [27,59–61] using fuzzy logic [111]. However, approaches like MergeSD [54] use overlapping intervals.
- A recent approach presented in [91], TargetCluster, discretises a continuous target variable before applying a subgroup discovery algorithm based on a clustering approach.
- In [107], a methodology for grouping different variables as a target variable was presented. This proposal is based on clustering to separately find clusters as values of the target variable.

5.3 Domain knowledge in subgroup discovery

Using domain knowledge in data mining methods can improve the quality of data mining results [94]. In subgroup discovery, it can help to focus the search on the interesting subgroups related to the target variable by restricting the search space.

Different approaches to include domain knowledge in subgroup discovery have been presented recently:

- In [87], the authors presented the *Semantic Subgroup Discovery* in which semantically annotated knowledge sources are used in the data mining process as background knowledge. In this process, the results obtained have a complex structure which allows the experts to see novel relationships in the data.
- In [7], *Domain Knowledge* is presented as a “methodological approach for providing domain knowledge in a declarative manner”.
- In [12,15], the authors introduce semi-automatic approaches using causal domain knowledge to improve the results of a subgroup discovery algorithm. In these studies, the domain knowledge is used to identify potential causal relations and confounding subgroup patterns. In addition, the use of supporting factors, defined as the values of several variables of the database that characterise a subgroup that can help to confirm that an individual is member of the subpopulation defined by the subgroup description [51], can be useful for the refinement of the extracted knowledge [11,14] using a case-based method for subgroup analysis and characterisation.

5.4 Filtering the rules obtained

Usually, data mining techniques return a set of rules too broad to be analysed by the experts. This is a general problem in the data mining techniques that also affects the task of discovery of subgroups. It is therefore sometimes necessary to reduce or filter the set of rules. In [23], a heuristic algorithm to reduce the number of rules obtained can be found.

For subgroup discovery, a proposal related to this task can be found:

- In [97], an iterative ranking approach to select the most important subgroups is presented, focusing on a subset of the database composed of certain attributes.

5.5 Visualisation of the interesting subgroups

The visualisation of results is very important for the usability of the knowledge extracted. Subgroup discovery algorithms are often used to present the results to an expert, who will take decisions based on these data.

- In [9], different visualisation methods supporting explorative and descriptive data mining were presented, implemented in the VIKAMINE system (see Appendix A.2).
- An interesting study including different representations of the subgroup discovery results available in the Orange data mining software (see Appendix A.1).
- An improvement for the visualisation of the interesting subgroups was presented in [69], with particular emphasis on the effects of the unusualness (13).
- In [62,89], a visual interactive subgroup discovery procedure which allows the navigation in the space of subgroups in a two-dimensional plot was shown. The authors used distribution rules for representing the knowledge.

5.6 Other approaches to subgroup discovery

Subgroup discovery has not only been presented in the specialised bibliography as a model to describe knowledge with respect to a target variable. Other approaches to subgroup discovery are listed below:

- In [117], an algorithm based on clustering for extracting subgroups is presented. This approach proposes using cluster-grouping [117] as a subtask to perform the subgroup discovery. The algorithm uses unusualness (13) as a quality function to prune the search space.
- An algorithm of subgroup discovery is used for discovering contrast sets in [77]. The idea is to solve the contrast set mining task by transforming it into a subgroup discovery task. This can be done using different techniques: one target value versus the rest (one versus all) or for every pair of values of the target variable (round robin). In the process, the algorithm uses unusualness (13) and several heuristics to prune the search space.
- In [19], the rule cubes technique is associated to the subgroup discovery concept. A rule cube is a discrete frequency distribution over a small set of variables, among which is the class attribute. The authors use similar quality measure to the used in subgroup discovery such as Piatetsky–Shapiro or statistical measures based on the contingency table. However, a causal analysis with respect to the rules to study the problem presented is performed.

6 Applications

A wide range of contributions in the specialised literature related to different fields can be found, where descriptive knowledge associated with a specific target value has a special interest. Table 7 summarises the real-world applications of algorithms of subgroup discovery.

In the following sections, different applications of subgroup discovery algorithms in different domains are shown together with the main features and properties used. For all of them, we present a short description together with a table containing the *Database* used, the *Algorithms* employed, and the number of instances (n_s) and variables (n_v) of the database.

Table 7 Real problems solved through subgroup discovery algorithms

Field	Application	References
Medical domain	Detection of risk groups with coronary heart disease	[43–45,48,49,80,84]
	Brain ischaemia	[47,52,76]
	Cervical cancer	[103]
	Psychiatric emergency	[29]
Bioinformatic domain	Leukaemia cancer	[50,104,105,115]
	Leukaemia$_{improved}$ cancer	[106]
	Subtypes of leukaemia ALL cancer	[106]
	Cancer diagnosis	[46,50,80,104–106,115]
	Interpreting positron emission tomography (PET) scans	[99]
Marketing	Financial	[71]
	Comercial	[44,84]
	Planning trade fairs	[18,61]
Learning	e-learning	[30,96]
Spatial subgroup mining	Demographic data	[6]
	Census data	[72,73,75]
	Vegetation data	[90]
Others	Traffic accidents	[64,65,67]
	Production control	[71]
	Mutagenesis	[86,113,114]
	Social data	[79]
	Voltage sag source location	[16]

Table 8 Description of the databases used in the medical domain

Database	Algorithm	n_s	n_v
Detection coronary heart disease [43–45,48,49,80,84]	SD	238	22
Brain ischaemia [47,52,76]	SD	300	26
Cervical cancer [103]	SD	Not esp.	Not esp.
Psychiatric emergency [29]	SDIGA, MESDIF, CN2-SD	925	72

If the number of instances or variables are not indicated in the problem, the corresponding columns are marked “Not esp.”, and multi-relational databases are marked “Multi-relational”.

6.1 Subgroup discovery in the medical domain

In the specialised literature, the following groups of problems addressed by subgroup discovery can be found: detection of risk groups with coronary heart disease, brain ischaemia, cervical cancer, and psychiatric emergency. In Table 8, the main properties of the databases used can be observed.

Almost all the proposals in the medical domain were solved through the DMS software[1] with the algorithm SD [43]. The psychiatric problem presented in [29] was solved with evolutionary fuzzy systems. These applications are listed below.

- In the detection of risk groups with coronary heart disease, the main goal was to obtain interesting subgroups for different problems (anamnestic information and physical examination results, results of laboratory tests, and electrocardiogram results). Initially, it was tackled in [43] and analysed by an expert. The problem has also been analysed in a number of contributions [44,45,48,49,80,84].
- In brain ischaemia, the goal was the extraction of useful knowledge which can help in diagnosis, prevention, and better understanding of the vascular brain disease. Different analysis was presented in [47,52].

  An analysis with SD, CN2-SD, and APRIORI-SD algorithms was performed in [76] to observe the differences among the subgroups obtained.
- The problem of cervical cancer was analysed in [103]. The objective was to establish the factors which influence cervical cancer diagnosis in a particular region of India. The process of extraction of interesting subgroups was guided by experts.
- A psychiatric emergency problem was presented in [29], where a comparison between SDIGA, MESDIF, and CN2-SD can be observed. The goal was to characterise subgroups of patients who tend to visit the psychiatric emergency department during a certain period of time. In this way, it is possible to improve the psychiatric emergency department organisation.

6.2 Bioinformatic problems solved through subgroup discovery

Different real problems have been solved using subgroup discovery algorithms in the bioinformatic domain. These problems are characterised by their high number of variables and low number of records in the databases. This makes it difficult to extract interesting results.

The real-world problems solved are: gene expressions data and interpreting positron emission tomography (PET) scans. The main properties of the databases used can be observed in Table 9.

Initially, the detection of subgroups in problems of cancer were tackled in [46,50,80] with the SD algorithm adding feature filtering methods. Then, in order to improve the interpretability of these results, the algorithm RSD was applied combined with other artificial intelligence techniques such as classifiers [115] and gene ontologies [104,105].

In [106], the authors used an improved version of a leukaemia database to obtain better results with the RSD algorithm. Furthermore, they searched subgroups of different subtypes of cancer for the value "ALL". The main goal in these papers was the obtention of gene expressions subgroups to detect diagnoses related to the cancer.

The interpretation of PET scans was analysed in [99] with the RSD algorithm. The goal was to describe the relationships between the scans and knowledge obtained from the database through data mining. First, the authors obtained the value of the target variable through clustering of the scans and later they applied RSD to obtain interesting subgroups.

6.3 Subgroup discovery in marketing

In this domain, a number of algorithms are applied to different real problems, which are described in Table 10.

[1] http://dms.irb.hr/index.php.

Table 9 Description of the databases used in the bioinformatic domain

Database	Algorithm	n_s	n_v
Leukaemia cancer [50,104,105,115]	SD,RSD	72	7.129
Leukaemia$_{improved}$ cancer [106]	RSD	73	6.817
Subtypes of leukaemia ALL cancer [106]	RSD	132	22.283
Cancer diagnosis [46,50,80,104–106,115]	SD,RSD	198	16.063
PET scans [99]	RSD	1.100	42

Table 10 Description of the databases used in the marketing domain

Database	Algorithm	n_s	n_v
Financial market in Germany [71]	EXPLORA	Not esp.	Not esp.
Recognise commercial brands [44]	SD	100	Not esp.
Direct mailing campaign [84]	CN2-SD	Multi-relational	
Natural non-alcoholic brand [84]	CN2-SD, supporting factors	Multi-relational	
Planning of trade fairs [18,61]	SDIGA,MESDIF	228	104

- A first application was briefly mentioned in [71] with the EXPLORA algorithm, where an analysis of financial market in Germany with data of different institutions was studied. In this problem, different preprocessing methods were necessary in order to obtain interesting subgroups. In the problem, general and large subgroups were obtained since the algorithm was used with disjunctions in the description language.
- An analysis of market with respect to determined brands was performed in [44] with the SD algorithm. However, in [84], other problems of the marketing campaign were studied with the algorithm CN2-SD and supporting factors. In both papers, the goal was to discover potential customers for different market brands. This problem was also tackled using the CN2 algorithm in [42].
- The planning of trade fairs was analysed with respect to evolutionary algorithms for subgroup discovery. The goal was to obtain subgroups for the planning of trade fairs which are considered by the businesses as an instrument for facilitating the attainment of commercial objectives.

6.4 Subgroup discovery in e-learning

The design of web-based education systems has had a high growth in the last years. These systems accumulate a great amount of valuable information when analysing students' behaviour or detecting possible errors, shortcomings, and improvements. However, due to the huge quantities of data, these systems can generate data mining tools which can assist in this task are demanded [95]. In [30,96], different algorithms of subgroup discovery and comparison among them with respect to systems based on electronic learning (e-learning) can be observed. For this problem, the data were obtained from the Moodle e-leaning system implanted in the University of Cordoba, and the main properties of the data can be observed in Table 11. The objective is to obtain knowledge from the usage data and to use it to improve the marks obtained by the students.

Table 11 Description of the databases used in the e-learning domain

Database	Algorithm	n_s	n_v
e-learning [30,96]	SDIGA,MESDIF,CN2-SD,APRIORI-SD	293	11

A first approach of the SDIGA algorithm (based on accuracy [96], coverage (1), and significance (9)) applied to this problem was used in [96], where a comparison study was presented. In [30], a different version of SDIGA was presented (based on unusualness (13), support (2), and confidence (3)) and compared with APRIORI-SD, CN2-SD, and MESDIF.

6.5 Spatial subgroup discovery

The combination of exploratory analysis of spatial data through geographical visualisation and techniques of subgroup discovery is discussed in different papers in the specialised bibliography. The problems tackled were related to census [72,73,75], demographic [6], and vegetation [90] data. The different proposals were used with different algorithms of subgroup discovery: SubgroupMiner, MIDOS, and Spatial Subgroup Mining, respectively.

The main property of these systems is the integration of the results obtained by different artificial intelligence techniques in a geographical information system. With this synergy, experts obtain a better interpretation and visualisation of the results, an important aspect in the subgroup discovery task.

- In [6], the MIDOS algorithm was used to checking whether different districts differ in thematic properties, in particular in population structure. In spite of the ability of the algorithm to use relational data, the database was transformed into one main table by the analyst in order to obtain the $Target_{value}$ and the properties to study.
- In [75], the SubgroupMiner algorithm was used on census data to search for possible effects on mortality. The objective was to analyse geographical factors, deprivation factors, transportation lines and so on.
- In [90], an approach focused on statistical significance (a measure similar to χ^2) was applied to vegetation data from Niger. The main goal was to evaluate subgroups related to properties of the zone through contingency tables.

6.6 Other applications

Other applications can be found in the specialised bibliography. The main properties of each problem solved by these applications of subgroup discovery algorithms can be observed in Table 12.

- In [65,67], a study related to traffic accidents was performed. These papers presented different comparisons such as CN2-SD versus SubgroupMiner. In [64], a comparison between SubgroupMiner, CN2-SD, and APRIORI-SD was presented.
- In [71], a production control problem was performed with the EXPLORA algorithm. The authors showed several main criteria for obtaining interesting subgroups, but as conclusions the authors mentioned the difficulty of obtaining the best hypothesis in the subgroup discovery task.
- The first application of the RSD algorithm was with *Mutagenesis* data (related to the traffic of calls in an enterprise). In [113], four versions of the RSD algorithm with different combinations of quality measures were tested, where the best results were obtained with

Table 12 Description of the databases

Database	Algorithm	n_s	n_v
Traffic accidents [64,65,67]	CN2-SD, APRIORI-SD, SubgroupMiner	Multi-relational	
Production control [71]	EXPLORA	Not esp.	Not esp.
Mutagenesis [86,113,114]	RSD	Not esp.	Not esp.
Social data [79]	SD	347	957
Voltage sag source location [16]	CN2-SD	430	Not esp.

the combination of unusualness (13) and weighting coverage [113]. Similar studies were presented in [86,114].

- In [79], an effort to analyse social data was performed with the SD algorithm. The authors studied different approaches related to social sciences, and therefore, different databases were constructed. Subgroup discovery applied to social data provides automatic detection of relevant information and descriptive information of the original data.
- In [16], an application of CN2-SD algorithm to voltage sag source location is proposed. The process is guided by an expert to find the more interesting subgroups and filtering them.

7 Conclusions

A survey of research on subgroup discovery has been provided in this paper, attempting to cover the early work in the field as well as recent literature related to the topic. The main properties and elements of this task have been presented, and the more widely used quality measures for subgroup discovery have been described and classified with respect to their properties. State-of-the-art and recent subgroup discovery algorithms have been briefly described and classified with respect to their foundations.

In addition, different applications of subgroup discovery approaches to real-world problems have been presented, organised with respect to the area of the application.

This is an emergent field, and there are several open problems in subgroup discovery. An important problem to address is to determine which quality measures are more adapted both to evaluating the subgroups discovered and to guiding the search process. A wide number of measures have been used, but there is no current consensus in the field about which are the most suitable measures for both processes. On the other hand, the discretisation of the continuous variables and its influence in the results of the subgroup discovery task is another open topic. It is unclear how the previous discretisation of continuous variables may affect the results of the subgroup discovery process, or the advantages of the subgroup discovery algorithms that use continuous variables without any prior discretisation. Another issue to be dealt with in more depth is the scalability of the subgroup discovery algorithms. Many of the subgroup discovery algorithms have a high computational cost when they are applied to large data sets. It would be interesting to develop scalable algorithms, or to determine, according to the features of the data set, a suitable reduced data set to apply the subgroup discovery algorithms. Finally, the combination between the subgroup discovery task and other fields such as semantic data, contrast set, clustering, and so on is beginning to be used in this area. This synergy helps to solve present problems in these fields like the use of unusualness in the contrast set. Furthermore, it improves the representation and comprehension of the results, for example through the semantic.

Acknowledgments This paper was supported by the Spanish Ministry of Education, Social Policy and Sports under projects TIN-2008-06681-C06-01 and TIN-2008-06681-C06-02 and by the Andalusian Research Plan under project TIC-3928.

Appendix A: Software tools with subgroup discovery algorithms

In this appendix the most used open source software tools implementing subgroup discovery algorithms are described. Next, for each tool, a subsection can be observed.

A.1 Orange

Orange is an open source tool for data visualisation and analysis, which can be downloaded in http://www.ailab.si/orange/. It can be installed using the most important operating systems such as Windows, Mac OS X and Linux.

Orange is a library of C++ core objects and routines that includes a large variety of standard and not-so-standard machine learning and data mining algorithms, plus routines for data input and manipulation. Orange is also a scriptable environment for fast prototyping of new algorithms and testing schemes. It is a collection of Python-based modules which are based on the core library and implement some functionality. A complete description of this software can be studied in [36].

A module for subgroup discovery in Orange is distributed free under GPL and can be downloaded from the website http://kt.ijs.si/petra_kralj/SubgroupDiscovery/. This tool implements three subgroup discovery algorithms—SD, CN2-SD, and Apriori-SD—and two visualisation methods—BAR and ROC—. This module permits the use of six different evaluation measures.

A.2 VIKAMINE software

VIKAMINE (Visual, Interactive, Knowledge-Intensive, (Data) Analysis, and Mining Environment) is an integrated system for visual analytics, data mining, and business intelligence. It features several powerful and intuitive visualisations complemented by fast automatic mining methods. It can be downloaded in the website http://vikamine.sourceforge.net/. The VIKAMINE software is released under the GNU Lesser General Public License (LGPL).

This software is based on Java, and the users can use it under Windows 95/98/ME/NT/2000/XP; for Mac OS X and Unix, the software is almost completely supported. A complete description can be studied in [9].

For the subgroup discovery task, the algorithm SD-Map is implemented in this tool. The user can choose between an automatic search of subgroup discovery, an interactive search of subgroups, or a combination of both. Afterwards, different comparisons, introspections, and analysis using the visualisations and options provided by the context menus in the subgroup results can be performed.

A.3 KEEL

KEEL (Knowledge Extraction based on Evolutionary Learning) is a open source (GPLv3) Java software tool which empowers the user to assess the behaviour of evolutionary learning and Soft Computing based techniques for different kinds of data mining problems like regression, classification, clustering, pattern mining, subgroup discovery, and so on. This software tool can be downloaded in the website http://www.keel.es.

The presently available version of KEEL consists of the following function blocks:

- Data Management, which is composed by a set of tools that can be used to build new data, export and import, edition, and so on.
- Design of Experiments whose aim is the design of the desired experimentation over the selected data, among them subgroup discovery.
- Design of Imbalanced Experiments whose aim is the design of the desired experimentation over the selected imbalanced data sets.
- Design of Experiments with vague data whose aim is the design of the desired experimentation over the selected data sets which contains a mixture of crisp and fuzzy values.
- Statistical tests. KEEL is one of the fewest Data Mining software tools that provides to the researcher a complete set of statistical procedures for pairwise and multiple comparisons.
- Education Experiments which allows to design an experiment which can be step-by-step debugged.

A complete description of this software tool can be studied in [4,5].

In the subgroup discovery module, the evolutionary algorithms presented up to the moment for this task can be used: SDIGA, MESDIF, and NMEEF-SD. This tool allows the user to set up different experiments for each algorithm. Furthermore, these algorithms can be combined with:

- Different preprocessing algorithms to improve the results obtained in the data set such as sampling, feature selection, and so on.
- Different visualisation methods like statistical tests to show their behaviour, or visualisations of the results obtained.

References

1. Abudawood T, Flach P (2009) Evaluation measures for multi-class subgroup discovery. In: Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases, vol 5781. Springer, LNAI, pp 35–50
2. Agrawal R, Imieliski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data. ACM Press, pp 207–216
3. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. AAAI Press, Cambridge, pp 307–328
4. Alcalá-Fdez J, Sánchez L, García S, del Jesus M, Ventura S, Garrell J, Otero J, Romero C, Bacardit J, Rivas V, Fernández J, Herrera F (2009) KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput 13(3):307–318
5. Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2010) KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Multiple Valued Logic Soft Comput (in press)
6. Andrienko N, Andrienko G, Savinov A, Voss H, Wettschereck D (2001) Exploratory analysis of spatial data using interactive maps and data mining. Cartogr Geogr Inf Sci 28(3):151–165
7. Atmueller M, Seipel D (2009) Using declarative specifications of domain knowledge for descriptive data mining. In: Proceedings of the international conference on applications of declarative programming and knowledge management and the workshop on logic programming, vol 5437. Springer, LNAI, pp 149–164
8. Atzmueller M, Lemmerich F (2009) Fast subgroup discovery for continuous target concepts. In: Proceedings of the 18th international symposium on methodologies for intelligent systems, vol 5722. Springer, LNAI, pp 35–44
9. Atzmueller M, Puppe F (2005) Semi-automatic visual subgroup mining using VIKAMINE. J Univers Comput Sci 11(11):1752–1765

10. Atzmueller M, Puppe F (2006) SD-Map—a fast algorithm for exhaustive subgroup discovery. In: Proceedings of the 17th European conference on machine learning and 10th European conference on principles and practice of knowledge discovery in databases, vol 4213. Springer, LNCS, pp 6–17
11. Atzmueller M, Puppe F (2008) A case-based approach for characterization and analysis of subgroup patterns. Appl Intell 28(3):210–221
12. Atzmueller M, Puppe F (2009) Knowledge discovery enhanced with semantic and social information, Springer, chap A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery, pp 19–36
13. Atzmueller M, Puppe F, Buscher HP (2004) Towards knowledge-intensive subgroup discovery. In: Proceedings of the Lernen-Wissensentdeckung-Adaptivität-Fachgruppe Maschinelles Lernen, pp 111–117
14. Atzmueller M, Baumeister J, Puppe F (2006) Introspective subgroup analysis for interactive knowledge refinement. In: Proceedings of the 9th international Florida artificial intelligence research society conference. AAAI Press, pp 402–407
15. Atzmueller M, Puppe F, Buscher HP (2009) A semi-automatic approach for confounding-aware subgroup discovery. Int J Artif Intell Tools 18(1):81–98
16. Barrera V, López B, Meléndez J, Sánchez J (2008) Voltage sag source location from extracted rules using subgroup discovery. Front Artif Intell Appl 184:225–235
17. Bay S, Pazzani M (2001) Detecting group differences: mining contrast sets. Data Mining Knowl Discov 5:213–246
18. Berlanga FJ, del Jesus MJ, González P, Herrera F, Mesonero M (2006) Multiobjective evolutionary induction of subgroup discovery fuzzy rules: a case study in marketing. In: Proceedings of the 6th industrial conference on data mining, vol 4065. Springer, LNCS, pp 337–349
19. Blumenstock A, Schweiggert F, Mueller M, Lanquillon C (2009) Rule cubes for casual investigations. Knowl Inf Syst 18(1):109–132
20. Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases, vol 5781. Springer, LNAI, pp 179–194
21. Box G, Jenkins G, Reinsel G (2008) Time series analysis: forecasting and control, 4th edn. Wiley, New York
22. Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data. ACM Press, pp 255–264
23. Bringmann B, Zimmermann A (2009) One in a million: picking the right patterns. Knowl Inf Syst 18(1):61–81
24. Cano JR, García S, Herrera F (2008) Subgroup discover in large size data sets preprocessed using stratified instance selection for increasing the presence of minority classes. Patt Recognit Lett 29:2156–2164
25. Cano JR, Herrera F, Lozano M, García S (2008) Making CN2-SD subgroup discovery algorithm scalable to large size data sets using instance selection. Expert Syst Appl 35:1949–1965
26. Carmona CJ, González P, del Jesus MJ, Herrera F (2009a) An analysis of evolutionary algorithms with different types of fuzzy rules in subgroup discovery. In: Proceedings of the IEEE international conference on fuzzy systems, pp 1706–1711
27. Carmona CJ, González P, del Jesus MJ, Herrera F (2009b) Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In: Proceedings of the 4th international conference on hybrid artificial intelligence systems, vol 5572. Springer, LNAI, pp 573–580
28. Carmona CJ, González P, del Jesus MJ, Herrera F (2010a) NMEEF-SD: Non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans Fuzzy Syst 18(5):958–970
29. Carmona CJ, González P, del Jesus MJ, Navío M, Jiménez L (2010b) Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. Soft Comput Special Issue on "Genetic Fuzzy Systems" (in press)
30. Carmona CJ, González P, del Jesus MJ, Romero C, Ventura S (2010c) Evolutionary algorithms for subgroup discovery applied to e-learning data. In: Proceedings of the IEEE international education engineering, pp 983–990
31. Cherkassky V, Mulier FM (2007) Learning from data: concepts, theory and methods, 2nd edn. IEEE Press, New York
32. Clark P, Boswell R (1991) Rule Induction with CN2: some recent improvements. In: Proceedings of the 5th European conference on machine learning, vol 482. Springer, LNCS, pp 151–163
33. Clark P, Niblett T (1989) The CN2 induction algorithm. Mach Learn 3:261–283
34. Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases. World Scientific, Singapore

35. Deb K, Pratap A, Agrawal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197
36. Demsar J, Zupan B, Leban G (2004) White Paper (http://www.ailabsi/orange)
37. Domingo C, Gavaldá R, Watanabe O (2002) Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining Knowl Discov 6(2):131–152
38. 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. ACM Press, pp 43–52
39. Dubois D, Prade H, Sudkamp T (2005) On the representation, measurement, and discovery of fuzzy associations. IEEE Trans Fuzzy Syst 13:250–262
40. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: 13th International joint conference on artificial intelligence, pp 1022–1029
41. Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 1–34
42. Flach PA, Gamberger D (2001) Subgroup evaluation and decision support for a direct mailing marketing problem. In: Proceedings of the 12th European conference on machine learning and 5th European conference on principles and practice of knowledge discovery in databases, pp 45–56
43. Gamberger D, Lavrac N (2002) Expert-guided subgroup discovery: methodology and application. J Artif Intell Res 17:501–527
44. Gamberger D, Lavrac N (2002) Generating actionable knowledge by expert-guided subgroup discovery. In: Proceedings of the 6th European conference on principles and practice of knowledge discovery in databases, vol 2431. Springer, LNCS, pp 163–174
45. Gamberger D, Lavrac N (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artif Intell Med 28(1):27–57
46. Gamberger D, Lavrac N (2004) Avoiding data overfitting in scientific discovery: experiments in functional genomics. In: Proceedings of the 16th European conference on artificial intelligence. IOS Press, pp 470–474
47. Gamberger D, Lavravc N (2007) Supporting factors in descriptive analysis of brain ischaemia. In: Proceedings of the 11th conference on artificial intelligence in medicine, vol 4594. Springer, LNCS, pp 155–159
48. Gamberger D, Lavrac N, Wettschereck D (2002) Subgroup visualization: a method and application to population screening. In: Proceedings of the 2nd international workshop on integration and collaboration aspects of data mining, decision support and meta-learning, pp 35–40
49. Gamberger D, Smuc T, Lavrac N (2003) Subgroup discovery: on-line data minig server and its application. In: Proceedings of the 5th international conference on simulations in biomedicine, pp 433–442
50. Gamberger D, Lavrac N, Zelezny F, Tolar J (2004) Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. J Biomed Inform 37(4):269–284
51. Gamberger D, Krstacic A, Krstatic G, Lavrac N, Sebag M (2005) Data analysis based on subgroup discovery: experiments in brain ischaemia domain. In: Proceedings of the 10th international workshop on intelligent data analysis in medicine and pharmacology, pp 52–56
52. Gamberger D, Lavrac N, Krstaic A, Krstaic G (2007) Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis. Appl Intell 27(3):205–217
53. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co, Reading
54. Grosskreutz H, Rueping S (2009) On subgroup discovery in numerical domains. Data Mining Knowl Discov 19(2):210–216
55. Grosskreutz H, Rueping S, Wrobel S (2008) Tight optimistic estimates for fast subgroup discovery. In: European conference on machine learning and principles and practice of knowledge discovery in databases, pp 440–456
56. Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data. ACM Press, pp 1–12
57. Herrera F (2008) Genetic fuzzy systems: taxomony, current research trends and prospects. Evol Intell 1:27–46
58. Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor
59. del Jesus MJ, González P, Herrera F (2007) Fuzzy sets and their extensions: representation, aggregation and models, vol 220, Springer, chap Subgroup Discovery with Linguistic Rules, pp 411–430
60. del Jesus MJ, González P, Herrera F (2007) Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: Proceedings of the IEEE symposium on computational intelligence in multicriteria decision making. IEEE Press, pp 50–57

61. del Jesus MJ, González P, Herrera F, Mesonero M (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans Fuzzy Syst 15(4):578–592
62. Jorge AM, Pereira F, Azevedo PJ (2006) Visual interactive subgroup discovery with numerical properties of interest. In: Proceedings of the 9th international conference on discovery science, vol 4265. Springer, LNAI, pp 301–305
63. Jovanoski V, Lavrac N (2001) Classification rule learning with APRIORI-C. In: 10th Portuguese conference on artificial intelligence on progress in artificial intelligence, knowledge extraction, multi-agent systems, logic programming and constraint solving, vol 2258. Springer, LNCS, pp 44–51
64. Kavsek B, Lavrac N (2004) Analysis of example weighting in subgroup discovery by comparison of three algorithms on a real-life data set. In: Proceedings of the 15th European conference on machine learning and 8th European conference on principles and practice of knowledge discovery in databases, pp 64–76
65. Kavsek B, Lavrac N (2004) Using subgroup discovery to analyze the UK traffic data. Metodoloski Zvezki 1(1):249–264
66. Kavsek B, Lavrac N (2006) APRIORI-SD: adapting association rule learning to subgroup discovery. Appl Artif Intell 20:543–583
67. Kavsek B, Lavrac N, Bullas JC (2002) Rule induction for subgroup discovery: a case study in mining UK traffic accident data. In: International multi-conference on information society, pp 127–130
68. Kavsek B, Lavrac N, Jovanoski V (2003) APRIORI-SD: adapting association rule learning to subgroup discovery. In: Proceedings of the 5th international symposium on intelligent data analysis, vol 2810. Springer, LNCS, pp 230–241
69. Kavsek B, Lavrac N, Todorovski L (2004) ROC analysis of example weighting in subgroup discovery. In: Proceedings of the 1st workshop on international workshop ROC analysis in artificial intelligence, pp 55–60
70. Kloesgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271
71. Kloesgen W (1999) Applications and research problems of subgroup mining. In: Proceedings of the 11th international symposium on foundations of intelligent systems. Springer, pp 1–15
72. Kloesgen W, May M (2002) Census data mining—an application. In: Proceedings of the 6th European conference on principles of data mining and knowledge discovery, pp 65–79
73. Kloesgen W, May M (2002) Spatial subgroup mining integrated in an object-relational spatial database. In: Proceedings of the 6th European conference on principles of data mining and knowledge discovery, pp 275–286
74. Kloesgen W, Zytkow J (2002) Handbook of data mining and knowledge discovery, Oxford
75. Kloesgen W, May M, Petch J (2003) Mining census data for spatial effects on mortality. Intell Data Anal 7:521–540
76. Kralj-Novak P, Lavrac N, Zupan B, Gamberger D (2005) Experimental comparison of three subgroup discovery algorithms: analysing brain ischemia data. In: Proceedings of the 8th international multiconference information society, pp 220–223
77. Kralj-Novak P, Lavrac N, Gamberger D, Krstacic A (2009) CSM-SD: methodology for contrast set mining through subgroup discovery. J Biomed Inform 42(1):113–122
78. Kralj-Novak P, Lavrac N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey of constrast set, emerging pateern and subgroup mining. J Mach Learn Res 10:377–403
79. Lambach D, Gamberger D (2008) Temporal analysis of political instability through descriptive subgroup discovery. Confl Manag Peace Sci 25:19–32
80. Lavrac N (2005) Subgroup discovery techniques and applications. In: Proceedings of the 9th Pacific-Asia conference on knowledge discovery and data mining, vol 3518. Springer, LNCS, pp 2–14
81. Lavrac N, Flach PA, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of the 9th international workshop on inductive logic programming, vol 1634. Springer, LNCS, pp 174–185
82. Lavrac N, Flach P, Kavsek B, Todorovski L (2002) Rule induction for subgroup discovery with CN2-SD. In: Proceedings of the 2nd international workshop on integration and collaboration aspects of data mining, decision support and meta-learning, pp 77–87
83. Lavrac N, Zelezny F, Flach PA (2003) RSD: relational subgroup discovery through first-order feature construction. In: Proceedings of the 12th international conference inductive logic programming, vol 2583. Springer, LNCS, pp 149–165
84. Lavrac N, Cestnik B, Gamberger D, Flach PA (2004) Decision support through subgroup discovery: three case studies and the lessons learned. Mach Learn 57(1–2):115–143
85. Lavrac N, Kavsek B, Flach PA, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5:153–188

86. Lavrac N, Zelezny F, Dzeroski S (2005) Local patterns: theory and practice of constraint-based relational subgroup discovery. In: International seminar on local pattern detection, vol 3539. Springer, LNCS, pp 71–88
87. Lavrac N, Kralj-Novak P, Mozetic I, Podpecan V, Motaln H, Petek M, Gruder K (2009) Semantic subgroup discovery: using ontologies in microarray data analysis. In: Proceedings of the 31st annual international conference of the IEEE engineering in medicine and biology society. IEEE Press, pp 5613–5616
88. Liu H, Hussain F, Tan C, Dash M (2002) Discretization: an enabling technique. Data mining Knowl Discov 6:393–423
89. Lucas JP, Jorge AP, Pereira F, Pernas AM, Machado AA (2007) A tool for interactive subgroup discovery using distribution rules. In: Proceedings of the 13th Portuguese conference on artificial intelligence, vol 4874. Springer, LNAI, pp 426–436
90. May M, Ragia L (2002) Spatial subgroup discovery applied to the analysis of vegetation data. In: Proceedings of the 4th international conference on practical aspects of knowledge management, vol 2569. Springer, LNCS, pp 49–61
91. Moreland K, Truemper K (2009) Discretization of target attributes for subgroup discovery. In: Proceedings of the 6th international conference machine learning and data mining in pattern recognition, vol 5632. Springer, LNAI, pp 44–52
92. Mueller M, Rosales R, Steck H, Krishnan S, Rao B, Kramer S (2009) Subgroup discovery for test selection: a novel approach and its application to breast cancer diagnosis. In: Proceedings of the 8th international symposium on intelligent data analysis, vol 5772. Springer, LNCS, pp 119–130
93. Noda E, Freitas AA, Lopes HS (1999) Discovering interesting prediction rules wih a genetic algorithm. IEEE Congr Evol Comput 2:1322–1329
94. Richardson M, Domingos P (2003) Learning with knowledge from multiple experts. In: Proceedings of the 20th international conference on machine learning. AAAI Press, pp 624–631
95. Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. Expert Syst Appl 33(1):135–146
96. Romero C, González P, Ventura S, del Jesus MJ, Herrera F (2009) Evolutionary algorithm for subgroup discovery in e-learning: a practical application using Moodle data. Expert Syst Appl 36:1632–1644
97. Rueping S (2009) Ranking interesting subgroups. In: Proceedings of the 26th international conference on machine learning, pp 913–920
98. Scheffer T, Wrobel S (2002) Finding the most interesting patterns in a database quickly by using sequential sampling. J Mach Learn Res 3:833–862
99. Schmidt J, Hapfelmeier A, Mueller M, Perneczky R, Kurz A, Drzezga A, Kramer S (2010) Interpreting PET scans by structured patient data: a data mining case study in dementia research. Knowl Inf Syst 24(1):149–170
100. Scholz M (2005) Knowledge-based sampling for subgroup discovery. In: International seminar on local pattern detection, vol 3539. Springer, LNAI, pp 171–189
101. Siebes A (1995) Data Surveying: foundations of an inductive query language. In: Proceedings of the 1st international conference on knowledge discovery and data mining. AAAI Press, pp 269–274
102. Bäck T, Fogel D, Michalewicz Z (1997) Handbook of evolutionary computation. Oxford University Press, New York
103. Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson
104. Trajkovski I, Zelezny F, Tolar J, Lavrac N (2006) Relational descriptive analysis of gene expression data. In: Proceedings of the 3rd starting artificial intelligence researchers. IOS Press, pp 184–195
105. Trajkovski I, Zelezny F, Tolar J, Lavrac N (2006) Relational subgroup discovery for descriptive analysis of microarray data. In: Proceedings of the 2nd international symposium in computational life sciences, vol 4216. Springer, LNCS, pp 86–96
106. Trajkovski I, Zelezny F, Lavrac N, Tolar J (2008) Learning relational descriptions of differentially expressed gene groups. IEEE Trans Syst Man Cybern C 38(1):16–25
107. Umek L, Zupan B, Toplak M, Morin A, Chauchat JH, Makovec G, Smrke D (2009) Subgroup discovery in data sets with multi-dimensional responses: a method and a case study in traumatology. In: Proceedings of the 12th conference on artificial intelligence in medicine, vol 5651. Springer, LNAI, pp 265–274
108. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European symposium on principles of data mining and knowledge discovery, vol 1263. Springer, LNAI, pp 78–87
109. Wrobel S (2001) Inductive logic programming for knowledge discovery in databases. Springer, chap Relational Data Mining, pp 74–101

110. Wu X, Kumar V, Ross-Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2009) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
111. Zadeh LA (1975) The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III. Inf Sci 8–9:199–249, 301–357, 43–80
112. Zelezny F, Lavrac N (2006) Propositionalization-based relational subgroup discovery with RSD. Machine Learning 62:33–63
113. Zelezny F, Lavrac N, Dzeroski S (2003) Constraint-based relational subgroup discovery. In: Proceedings of the 2nd workshop on multi-relational data mining, pp 135–150
114. Zelezny F, Lavrac N, Dzeroski S (2003) Using constraints in relational subgroup discovery. In: International conference on methodology and statistics, pp 78–81
115. Zelezny F, Tolar J, Lavrac N, Stepankova O (2005) Relational subgroup discovery for gene expression data mining. In: Proceedings of the 3rd European medical and biological engineering conference
116. Zembowicz R, Zytkow JM (1996) From contingency tables to various forms of knowledge in databases. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 329–349
117. Zimmerman A, de Raedt L (2009) Cluster-grouping: from subgroup discovery to clustering. Mach Learn 77(1):125–159
118. Zitzler E, Laumanns M, Thiele L (2002) SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In: International congress on evolutionary methods for design optimization and control with applications to industrial problems, pp 95–100

Author Biographies

Francisco Herrera received the M.Sc. degree in Mathematics in 1988 and the Ph.D. degree in Mathematics in 1991, both from the University of Granada, Spain. He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has published more than 150 papers in international journals. He is coauthor of the book "*Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*" (World Scientific, 2001). As edited activities, he has co-edited five international books and co-edited twenty special issues in international journals on different Soft Computing topics. He acts as associated editor of the journals: IEEE Transactions on Fuzzy Systems, Information Sciences, Mathware and Soft Computing, Advances in Fuzzy Systems, Advances in Computational Sciences and Technology, and International Journal of Applied Metaheuristics Computing. He currently serves as area editor of the Journal Soft Computing (area of genetic algorithms and genetic fuzzy systems), and he serves as member of several journal editorial boards, among others: Fuzzy Sets and Systems, Applied Intelligence, Knowledge and Information Systems, Information Fusion, Evolutionary Intelligence, International Journal of Hybrid Intelligent Systems, Memetic Computation. His current research interests include computing with words and decision making, bibliometrics, data mining, data preparation, instance selection, fuzzy rule–based systems, genetic fuzzy systems, knowledge extraction based on evolutionary algorithms, memetic algorithms and genetic algorithms.

Cristóbal José Carmona received the M.Sc. degree in computer science from the University of Jaen, Spain, in 2006. He is a research cooperator in the Department of Computer Science, University of Jaen, Spain. Currently, he is working with Intelligent Systems and Data Mining Research Group of Jaen. His research interest includes subgroup discovery, contrast set mining, genetic fuzzy systems, genetic algorithm and data mining.

Pedro González received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 1991 and 2008, respectively. Currently, he is an Associate Professor with the Department of Computer Science, University of Jaen, Spain. His research interest include fuzzy rule-based systems, machine learning, genetic fuzzy systems, subgroup discovery, and data mining.

María José del Jesus received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 1994 and 1999, respectively. She is an Associate Professor with the Department of Computer Science, University of Jaen, Spain. Her research interests include fuzzy rule-based systems, fuzzy and linguistic modelling, fuzzy classification, machine learning, genetic fuzzy systems, and data mining.