Supervised Descriptive Rule Induction

Petra Kralj Novak1, Nada Lavrač1,2, and Geoffrey I. Webb3
1Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
2University of Nova Gorica, Nova Gorica, Slovenia
3Faculty of Information Technology, Monash University, Clayton, VIC, Australia

Synonyms

SDRI

Definition

Supervised descriptive rule induction (SDRI) is a machine learning task in which individual patterns in the form of rules (see classification rule) intended for interpretation are induced from data, labeled by a predefined property of interest. In contrast to standard supervised rule induction, which aims at learning a set of rules defining a classification/prediction model, the goal of SDRI is to induce individual descriptive patterns. In this respect, SDRI is similar to association rule discovery, but the consequents of the rules are restricted to a single variable – the property of interest – and, except for the discrete target attribute, the data is not necessarily assumed to be discrete.

Supervised descriptive rule induction assumes a set of training examples, described by attributes and their values and a selected attribute of interest (called the target attribute). Supervised descriptive rule induction induces rules that may each be interpreted independently of the others. Each rule is a local model, covering a subset of training examples, that captures a local relationship between the target attribute and the other attributes.

Induced descriptive rules are mainly aimed at human interpretation. More specifically, the purposes of supervised descriptive rule induction are to allow the user to gain insights into the data domain and to better understand the phenomena underlying the data.

Motivation and Background

Symbolic data analysis techniques aim at discovering comprehensible patterns or models in data. They can be divided into techniques for predictive induction, where models, typically induced from class-labeled data, are used to predict the class value of previously unseen examples, and descriptive induction, where the aim is to find comprehensible patterns, typically induced from unlabeled data. Until recently, these techniques have been investigated by two different research communities: predictive induction mainly by the machine learning community and descriptive induction mainly by the data mining community.
Data mining tasks where the goal is to find comprehensible patterns from labeled data have been addressed by both the machine learning and the data mining community independently. The data mining community, using the association rule learning perspective, adapted association rule learners like Apriori (Agrawal et al. 1996) to perform tasks on labeled data, like class association rule learning (Liu et al. 1998; Fürnkranz et al. 2012), as well as contrast set mining (Bay and Pazzani 2001) and emerging pattern mining (Dong and Li 1999). On the other hand, the machine learning community, which traditionally focused on the induction of rule sets from labeled data for the purposes of classification, turned to building individual rules for exploratory data analysis and interpretation. This is the goal of the task named subgroup discovery (Wrobel 1997). These are the main areas of supervised descriptive rule induction. All deal with finding comprehensible rules from class-labeled data. However, the methods used and the interpretation of the results differ slightly from approach to approach. Other related approaches include change mining, mining of closed sets for labeled data, exception rule mining, bump hunting, quantitative association rules, and impact rules. See Kralj Novak et al. (2009) for a more detailed survey of supervised descriptive rule induction.

**Structure of the Learning System**

Supervised descriptive rule induction assumes that there is data with the property of interest defined by the user. Let us illustrate supervised descriptive rule induction using data from Table 1, a very small artificial sample data set, adapted from Ross Quinlan (1986), which contains the results of a survey on 14 individuals, concerning the approval or disproval of an issue analyzed in the survey. Each individual is characterized by four attributes that encode rudimentary information about the sociodemographic background. The last column (Approved) is the designated property of interest, encoding whether the individual approved or disproved the issue. Unlike predictive induction, where the aim is to find a predictive model, the goal of supervised descriptive rule induction is to find local patterns in the form of individual rules describing individuals that are likely to approve or disprove the issue, based on the four demographic characteristics.

<table>
<thead>
<tr>
<th>Supervised Descriptive Rule Induction, Table 1</th>
<th>A sample database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>Marital status</td>
</tr>
<tr>
<td>Primary</td>
<td>Single</td>
</tr>
<tr>
<td>Primary</td>
<td>Single</td>
</tr>
<tr>
<td>Primary</td>
<td>Married</td>
</tr>
<tr>
<td>University</td>
<td>Divorced</td>
</tr>
<tr>
<td>University</td>
<td>Married</td>
</tr>
<tr>
<td>Secondary</td>
<td>Single</td>
</tr>
<tr>
<td>Secondary</td>
<td>Divorced</td>
</tr>
<tr>
<td>Secondary</td>
<td>Single</td>
</tr>
<tr>
<td>Secondary</td>
<td>Married</td>
</tr>
<tr>
<td>Primary</td>
<td>Married</td>
</tr>
<tr>
<td>Secondary</td>
<td>Divorced</td>
</tr>
<tr>
<td>University</td>
<td>Divorced</td>
</tr>
<tr>
<td>Secondary</td>
<td>Divorced</td>
</tr>
</tbody>
</table>
MaritalStatus=single AND Sex=male → Approved=no
Sex=male → Approved=no
Sex=female → Approved=yes
MaritalStatus=married → Approved=yes
MaritalStatus=divorced AND HasChildren=yes → Approved=no
MaritalStatus=single Approved=no

Supervised Descriptive Rule Induction, Fig. 1 Selected descriptive rules, describing individual patterns in the data of Table 1

while some descriptive approaches make no attempt at completeness, as they assess each pattern on its individual merits.

Exactly which rules will be induced by a supervised descriptive rule induction algorithm depends on the task definition, the selected algorithm, as well as the user-defined constraints concerning minimal rule support, precision, etc. Different learning approaches and heuristics have been proposed to induce supervised descriptive rules.

Applications

Applications of supervised descriptive rule induction are widely spread. See Kralj Novak et al. (2009) for a detailed survey.

Subgroup discovery has been used in numerous real-life applications Herrera et al. (2011). Medical applications include the analysis of coronary heart disease, brain ischemia data analysis, the analysis of cervical cancer, and psychiatric emergency, as well as profiling examiners for sonographic examinations. Spatial subgroup mining applications include mining of census data, mining of vegetation data and mining of demographic data. There are also applications in marketing, traffic accidents, production control, election analysis, and social data.

Contrast set mining has been used with retail sales data and for designing customized insurance programs. It has also been used in medical applications to identify patterns in synchrotron x-ray data that distinguish tissue samples of different forms of cancerous tumor and for distinguishing between groups of brain ischemia patients.

Emerging pattern mining has been mainly applied to the field of bioinformatics, more specifically to microarray data analysis. For example, an interpretable classifier based on simple rules that is competitive to the state of the art black-box classifiers on the acute lymphoblastic leukemia (ALL) microarray data set was built from emerging patterns. Another application was about finding groups of genes by emerging patterns in a ALL/AML data set and a colon tumor data set. Emerging patterns were also used together with the unexpected change approach and the added/perished rule to mine customer behavior.

Future Directions

A direction for further research is to decompose SDRI algorithms and preprocessing and evaluation methods into basic components and to reimplement them as connectable web services, which include the definition of interfaces between SDRI services. For instance, this can include the adaptation and implementation of subgroup discovery techniques to solving open problems in the area of contrast set mining and emerging patterns. This would allow for the improvement of algorithms due to the cross-fertilization of ideas from different SDRI subareas.

Another direction for further research concerns complex data types and the use of background knowledge. The SDRI attempts in this direction include relational subgroup discovery approaches like algorithms Midos (Wrobel 2001), RSD (relational subgroup discovery) (Železný and Lavrač 2006), and SubgroupMiner (Klösgen and May 2002), which is designed for spatial
data mining in relational space databases. When ontologies are used as background knowledge to define the hypothesis search space and data are used to constrain and guide the hypothesis search and evolution, and this is called semantic subgroup discovery (Vavpetič and Lavrač 2013).

Cross-References

- Apriori
- Association Rule Discovery
- Classification Rule
- Contrast Set Mining
- Emerging Pattern Mining
- Local Model
- Models
- Rule Sets
- Subgroup Discovery
- Supervised Rule Induction

Recommended Reading