Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A framework for redescription set construction

Matej Mihelčić[a,c,*], Sašo Džeroski[b,c], Nada Lavrač[b,c], Tomislav Šmuc[a]

[a] Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[b] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[c] International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 25 June 2016
Revised 6 October 2016
Accepted 6 October 2016
Available online 8 October 2016

Keywords:
Knowledge discovery
Redescription mining
Predictive clustering trees
Redescription set construction
Scalarization
Conjunctive refinement
Redescription variability

ABSTRACT

Redescription mining is a field of knowledge discovery that aims at finding different descriptions of similar subsets of instances in the data. These descriptions are represented as rules inferred from one or more disjoint sets of attributes, called views. As such, they support knowledge discovery process and help domain experts in formulating new hypotheses or constructing new knowledge bases and decision support systems. In contrast to previous approaches that typically create one smaller set of redescriptions satisfying a pre-defined set of constraints, we introduce a framework that creates large and heterogeneous redescription set from which user/expert can extract compact sets of differing properties, according to its own preferences. Construction of large and heterogeneous redescription set relies on CLUS-RM algorithm and a novel, conjunctive refinement procedure that facilitates generation of larger and more accurate redescription sets. The work also introduces the variability of redescription accuracy when missing values are present in the data, which significantly extends applicability of the method. Crucial part of the framework is the redescription set extraction based on heuristic multi-objective optimization procedure that allows user to define importance levels towards one or more redescription quality criteria. We provide both theoretical and empirical comparison of the novel framework against current state of the art redescription mining algorithms and show that it represents more efficient and versatile approach for mining redescriptions from data.

1. Introduction

In many scientific fields, there is a growing need to understand measured or observed data, to find different regularities or anomalies, groups of instances (patterns) for which they occur and their descriptions in order to get an insight into the underlying phenomena.

This is addressed by redescription mining (Ramakrishnan, Kumar, Mishra, Potts, & Helm, 2004), a type of knowledge discovery that aims to find different descriptions of similar sets of instances by using one, or more disjoint sets of descriptive attributes, called views. It is applicable in a variety of scientific fields like biology, economy, pharmacy, ecology, social science and other, where it is important to understand connections between different descriptors and to find regularities that are valid for different subsets of instances. Redescriptions are tuples of logical formulas which are called queries. Redescription $R_{ex} = (q_1, q_2)$ contains two queries:

q_1: $(-1.8 \leq \tilde{t}_7 \leq 4.4 \wedge 12.1 \leq \tilde{p}_6 \leq 21.2)$
q_2: Polarbear

The first query (q_1') describes a set of instances (geospatial locations) by using a set of attributes related to temperature (t) and precipitation (p) in a given month as first view (in the example average temperature in July and average precipitation in June). The second query (q_2') describes very similar set of locations by using a set of attributes specifying animal species inhabiting these locations as a second view (in this instance polar bear). Queries contain only conjunction logical operator, though the approach supports conjunction, negation and disjunction operators.

We first describe the fields of data mining and knowledge discovery closely related to redescription mining. Next, we describe recent research in redescription mining, relevant to the approach we propose. We then outline our approach positioned in the context of related work.

1.1. Fields related to redescription mining

Redescription mining is related to association rule mining (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996; Hipp, Güntzer, & Nakhaeizadeh, 2000; Zhang & He, 2010), two-view data

[*] Corresponding author.
E-mail addresses: matej.mihelcic@irb.hr (M. Mihelčić), saso.dzeroski@ijs.si (S. Džeroski), nada.lavrac@ijs.si (N. Lavrač), tomislav.smuc@irb.hr (T. Šmuc).

http://dx.doi.org/10.1016/j.eswa.2016.10.012
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

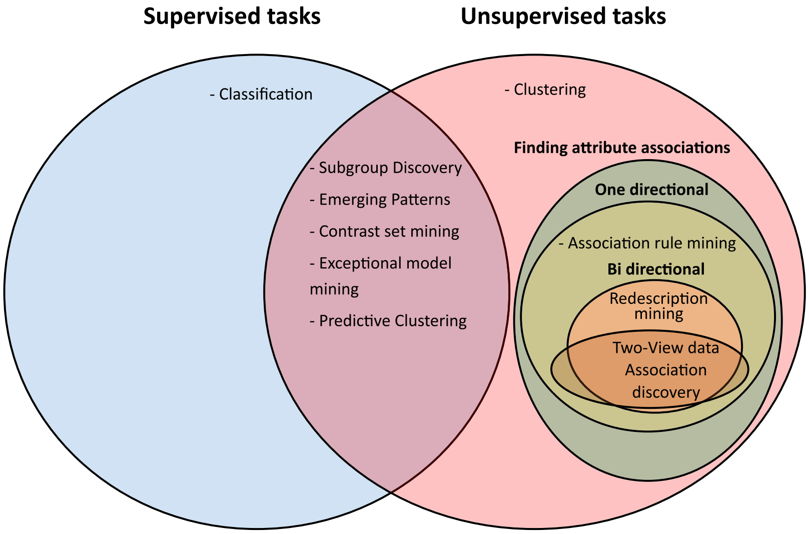

Fig. 1. Relation between redescription mining and other related tasks.

association discovery (van Leeuwen & Galbrun, 2015), clustering (Cox, 1957; Fisher, 1958; Jain, Murty, & Flynn, 1999; Ward, 1963; Xu & Tian, 2015) and it's special form conceptual clustering (Fisher, 1987; Michalski, 1980), subgroup discovery (Herrera, Carmona, González, & Jesus, 2010; Klösgen, 1996; Novak, Lavrač, & Webb, 2009; Wrobel, 1997), emerging patterns (Dong & Li, 1999; Novak et al., 2009), contrast set mining (Bay & Pazzani, 2001; Novak et al., 2009) and exceptional model mining (Leman, Feelders, & Knobbe, 2008). Most important relations can be seen in Fig. 1.

Association rule mining (Agrawal et al., 1996) is related to redescription mining in the aim to find queries describing similar sets of instances which reveal associations between attributes used in these queries. The main difference is that association rules produce one directional associations while redescription mining produces bi directional associations. Two-view data association discovery (van Leeuwen & Galbrun, 2015) aims at finding a small, non - redundant set of associations that provide insight in how two views are related. Produced associations are both uni and bi directional as opposed to redescription mining that only produces bi directional connections providing interesting descriptions of instances.

The main goal of clustering is to find groups of similar instances with respect to a set of attributes. However, it does not provide understandable and concise descriptions of these groups which are often complex and hard to find. This is resolved in conceptual clustering (Fisher, 1987; Michalski, 1980) that finds clusters and concepts that describe them. Redescription mining shares this aim but requires each discovered cluster to be described by at least two concepts. Clustering is extended by multi-view (Bickel & Scheffer, 2004; Wang, Nie, & Huang, 2013) and multi-layer clustering (Gamberger, Mihelčić, & Lavrač, 2014) to find groups of instances that are strongly connected across multiple views.

Subgroup discovery (Klösgen, 1996; Wrobel, 1997) differs from redescription mining in its goals. It finds queries describing groups of instances having unusual and interesting statistical properties on their target variable which are often unavailable in purely descriptive tasks. Exceptional model mining (Leman et al., 2008) extends subgroup discovery to more complex target concepts searching for subgroups such that a model trained on this subgroup is exceptional based on some property.

Emerging Patterns (Dong & Li, 1999) aim at finding itemsets that are statistically dependent on a specific target class while Contrast Set Mining (Bay & Pazzani, 2001) identifies monotone conjunctive queries that best discriminate between instances containing one target class from all other instances.

1.2. Related work in redescription mining

The field of redescription mining was introduced by Ramakrishnan et al. (2004), who present an algorithm to mine redescriptions based on decision trees, called CARTwheels. The algorithm works by building two decision trees (one for each view) that are joined in the leaves. Redescriptions are found by examining the paths from the root node of the first tree to the root node of the second. The algorithm uses multi class classification to guide the search between the two views. Other approaches to mine redescriptions include the one proposed by Zaki and Ramakrishnan (2005), which uses a lattice of closed descriptor sets to find redescriptions; the algorithm for mining exact and approximate redescriptions by Parida and Ramakrishnan (2005) that uses relaxation lattice, and the greedy and the MID algorithm based on frequent itemset mining by Gallo, Miettinen, and Mannila (2008). All these approaches work only on Boolean data.

Galbrun and Miettinen (2012b) extend the greedy approach by Gallo et al. (2008) to work on numerical data. Redescription mining was extended by Galbrun and Kimmig (2013) to a relational and by Galbrun and Miettinen (2012a) to an interactive setting. Recently, two tree-based algorithms have been proposed by Zinchenko (2014), which explore the use of decision trees in a non-Boolean setting and present different methods of layer-by-layer tree construction, which make informed splits at each level of the tree. Mihelčić, Džeroski, Lavrač, and Šmuc (2015a, b) proposed a redescription mining algorithm based on multi-target predictive clustering trees (PCTs) (Blockeel & De Raedt, 1998; Kocev, Vens, Struyf, & Džeroski, 2013). This algorithm typically creates a large number of redescriptions by executing PCTs iteratively: it uses rules created for one view of attributes in one iteration, as target attributes for generating rules for the other view of attributes in the next iteration. A redescription set of a given size is improved over the iterations by introducing more suitable redescriptions which replace the ones that are inferior according to predefined quality criteria.

In this work, we introduce a redescription mining framework that allows creating multiple redescription sets of user defined size, based on user defined importance levels of one or more redescription quality criteria. The underlying redescription mining