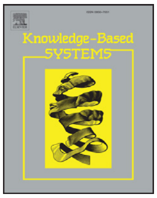
Knowledge-Based Systems 000 (2017) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy

C.J. Carmona[a,b,*], M.J. del Jesus[c], F. Herrera[d,e]

[a] Department of Civil Engineering, Languages and Computer Technology Systems, University of Burgos, Burgos, 09006, Spain
[b] Leicester School of Pharmacy, De Montfort University LE1 9BH, Leicester, United Kingdom
[c] Department of Computer Science, University of Jaen, Jaen, 23071, Spain
[d] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, 18001, Spain
[e] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 1 February 2017
Revised 11 October 2017
Accepted 12 October 2017
Available online xxx

Keywords:
Supervised descriptive rule discovery
Subgroup discovery
Emerging patterns
Contrast sets
Weighted relative accuracy

ABSTRACT

Supervised descriptive rule discovery represents a set of data mining techniques whose objective is to describe data with respect to a property of interest. This concept encompasses different techniques such as subgroup discovery, emerging patterns and contrast sets. Supervised learning is used to obtain rules for descriptive purposes but with different quality measures. Although their origin is based on different data mining tasks, our hypothesis is about the existence of a compatibility between subgroup discovery, emerging patterns and contrast sets thanks to the common use of a weighted relative accuracy quality measure. A complete analysis shows this relationship between the different tasks. The analysis is supported by an empirical study with the most representative algorithms for each technique.

The paper shows how the use of the weighted relative accuracy allows the experts to distinguish between interesting subgroups, emerging and/or contrasting rules thanks to the relation between the quality measures employed in the search process for different models. In addition, this relationship enables us to analyse the main differences and/or similarities between the different techniques within supervised descriptive rule discovery. This scenario opens up new challenges for the supervised descriptive rule learning models in analysing and developing descriptive models with a new perspective.

1. Introduction

Data mining is a computational process for discovering knowledge in data through the use of different methodologies, technologies and systems [15]. There are two areas clearly differentiated within data mining: predictive data mining, whose objective is to make predictions about future or unknown objects; and descriptive data mining, where the search for relationships between features in data is desired. In general, predictive induction is employed with supervised learning that assumes that objects are labelled and whose objective is to extract knowledge in order to predict values from one variable of interest. On the other hand, descriptive induction uses unsupervised learning combined with the analysis of unclassified objects. However, there is a group of techniques called Supervised Descriptive Rule Discovery (SDRD), defined in [28], where the main proposal is the search for interesting descriptions in data with respect to a property or class of interest. Essentially, SDRD describes labelled data.

The most representative techniques within SDRD are Subgroup Discovery (SD) [27,37], Emerging Patterns (EPs) [11] and Contrast Sets (CSs) [3]. All of these have been defined at different stages by different authors. Whereas the main idea in CSs is to search for contrasting relations among variables with respect to different groups, EPs describe emerging tendencies in data with respect to a time variable or search for distinct features with respect to a property of interest. On the other hand, SD describes interesting and unusual relationships in data with respect to a property of interest. In summary, their main goals are very similar and it is primarily the terminology that differs as well as the quality measures used in order to analyse a given problem.

Nowadays, the problematic within SDRD is that there is no consensus about the use of one or another quality measure in order to analyse the relevance of the proposals. In fact, there are a large number of quality measures for each technique and this complicates both the analysis of the knowledge extracted and guiding

[*] Corresponding author at: EPS Vena (A1-4127), University of Burgos, 09006, Burgos, Spain

E-mail addresses: cjcarmona@ubu.es, ccarmona@ujaen.es (C.J. Carmona), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.knosys.2017.10.015
0950-7051/© 2017 Elsevier B.V. All rights reserved.

the search process for the experts. Therefore, the knowledge extracted for a problem could be analysed from different perspectives in function of the quality measures employed.

This paper analyses the different data mining techniques within SDRD, their features, the type of knowledge extracted and the main quality measures considered. Specifically, we have the hypothesis concerning the existence of a common nexus among SD, EPs and CSs related to the weighted relative accuracy (*WRAcc*) quality measure (also known as unusualness) that measures the relationship between coverage and gain accuracy.

This contribution presents the *WRAcc* as the central axis in the analysis of SD, EPs and/or CSs. With the *WRAcc* value we could determine whether a rule is an interesting subgroup, EP and/or CS. It is important to highlight that this study shows the main behaviour for each technique within SDRD, allowing the experts to position SD, EPs and/or CSs with their main differences and/or similarities and so improve future studies.

Some interesting conclusions will be discussed as lessons learnt:

- The EP task attempts to obtain very precise rules regard less of the number of positive examples covered.
- In the CS task the main objective is the obtaining of rules with a high number of examples covered.
- For the SD task, the central axis is focused on the gain accuracy through the use of the *WRAcc*.

These lessons will also allow us to discuss some challenges in the topic. Specifically, this paper sets new foundations for the development and/or analysis of proposals within SDRD where the main axis in the analysis would be performed through the *WRAcc* quality measure. This paper opens up to the possible extension of the SDRD concept in order to include all those tasks with the same objective such as discriminative patterns [13] or change mining [32], for example. Finally, it is important to remark the skills of the SDRD techniques for the analysis of complex problems such as big data, unbalanced data, streaming data, etc. Through the use of the *WRAcc* quality measure in the different approaches studied this analysis would be simplified.

To do so, this paper is structured as follows: First, Section 2 provides a complete introduction to SDRD, showing the definition, main properties and state of the art for SD, EPs, and CSs. Section 3 introduces the compatibility of terms between the tasks included in SDRD and *WRAcc* as the key factor determining the connection between them. Next, Section 4 presents the empirical study based on the previous compatibilities where a complete analysis with different datasets and the most relevant algorithms for each technique is performed. Section 5 discussions challenges within SDRD tasks. Finally, the paper is concluded with the main findings in Section 6.

2. Supervised descriptive rule discovery

In data mining there are two main approaches used in order to analyse data: supervised learning (labelled data) and unsupervised learning (unlabelled data). Together with these approaches we further distinguish between predictive and descriptive induction, whereby predictive data mining methods are usually supervised (induce models from labelled data), and descriptive data mining methods are typically unsupervised (induce interesting association in unlabelled data).

The SDRD concept was introduced by Kralj-Novak et al. [28] in 2009. It describes the group of rule based techniques used in order to obtain descriptive knowledge with respect to labelled data. All techniques represented in this concept have as their objective the understanding of underlying phenomena instead of the classification of new instances.

An illustrative example for an SDRD model:

A medical center wants to know in what circumstances a patient may suffer a certain type of cancer; the intention is not to predict cancer, but to understand the risk factors that lead to this.

In Fig. 1 examples of the predictive supervised, descriptive unsupervised and SDRD tasks are presented in order to show the main differences and properties of the tasks included in the SDRD concept:

- Fig. 1(a) represents graphically the model obtained by a predictive algorithm based on the extraction of rules for classification. As can be observed, six rules (areas between dotted lines) divide the space into different areas that allow analysis of the problem in an easy way. In this way, the model is able to classify new instances of the problem with good values of precision.
- The model presented in Fig. 1(b) is an unsupervised descriptive model, e.g. clustering that groups unlabelled instances in different areas (circles). As can be observed, the model obtains three groups of instances with a soft overlapping between the lower and the remaining groups, with a simple and single interpretation for each group.
- On the other hand, Fig. 1(c) presents an SDRD model, where two rules (circles) for each value of the target variable are obtained. Rules are usually represented in a similar way to Fig. 1(a). Another important property is that the knowledge for each rule is considered as individual knowledge instead of rules dependant on one another. There is a possibility of overlapping between rules, as can be observed in the rules for the blue target value.

Throughout the literature the main models within SDRD have been classified in three different groups: SD, EPs and CSs. Next, the definitions and main properties are outlined for SD (Section 2.1), EPs (Section 2.2) and CSs (Section 2.3).

2.1. Subgroup Discovery

The SD was introduced by Kloesgen [27] and Wrobel [37] in 1996 and 1997, respectively. Its objective is to discover interesting relationships between different objects in a set with respect to a property of interest, widely known throughout the literature as class or target variable. The patterns extracted (called subgroups by Siebes [36]) are normally represented through rules [18], such as:

$$R = Class \leftarrow Cond$$

where *Cond* is a conjunction of attribute-value pairs and *Class* the property of interest. The examples containing the specific value for the *Class* are the positive examples and the remaining ($\overline{Class}$) the negative ones.

There is no consensus within SD about the use of a concrete quality measure, however the weighted accuracy relative (*WRAcc*) is the one most employed in the literature. This quality measure was defined as [30]:

$$WRAcc(Class \leftarrow Cond) = p(Cond) \cdot (p(Class|Cond) - p(Class)) \quad (1)$$

where a balance between generality, precision and gain accuracy is considered. The importance of this quality measure within the SDRD models is reflected in Section 3.2.

From the inception of the SD concept in 1996 there has been widespread application, especially in the last decades with the