Expert Systems With Applications 93 (2018) 223–231

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A novel characterisation-based algorithm to discover new knowledge from classification datasets without use of support

Enrique Lazcorreta Puigmartí[a,*], Federico Botella[a], Antonio Fernández-Caballero[b]

[a] Instituto Universitario Centro de Investigación Operativa (CIO), Universidad Miguel Hernández de Elche, 03202-Elche, Spain
[b] Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, 02071-Albacete, Spain

ARTICLE INFO

Article history:
Received 3 April 2017
Revised 11 October 2017
Accepted 11 October 2017
Available online 13 October 2017

Keywords:
Classification dataset
Catalogue
Association rules mining
Classification association rules mining

ABSTRACT

This paper introduces a novel proposal to discover the best associative classification rules through studying the influence of the attributes used in robust catalogues. Notice that a catalogue is defined as a dataset free of duplicate records. Moreover, a robust catalogue is obtained when incomplete records and those with uncertainty are eliminated from a catalogue. Therefore, a robust catalogue is a collection of association rules with 100% confidence and unitary support. In this paper we demonstrate that robust catalogues contain the same association rules as the datasets from which they were obtained, but can be managed in memory without eliminating any data from the analysis. In fact, the experiments performed show that all robust catalogues contained in a classification dataset are easily obtained, providing millions of associative classification rules with 100% confidence to the expert researcher in data mining.

1. Introduction

Data mining (Agrawal, Imielinski, & Swami, 1993) aims to efficiently discover patterns from large databases. Liu, Hsu, and Ma (1998) proposed the use of Association Rule Mining (ARM) techniques to form accurate classifiers, interpreting Association Rules (ARs) as Classification Association Rules (CARs). ARs are relations such as “*when items A and B belong to a sample, then item Calso belongs to the sample*”, generally “*antecedent* ⇒ *consequent*”. Unlike transactional databases, instances from classification datasets (datasets from now on) are ordered vectors with one value for each attribute and another value for the class. A CAR is a special case of association rule in which only the class attribute appears in the rule’s consequent. CARs are a subset of the ARs contained in a dataset.

In order to measure the quality of the ARs, *support* and *confidence* metrics are used. Support is the frequency of a rule in the dataset and confidence represents how many times the antecedent belongs to the samples having the rule consequent. A dataset has an exponential number of ARs, depending on the quantity of items and relations existing among their items. Therefore, large datasets are not completely analysed due to the fact that structures used in ARM algorithms require huge amounts of RAM memory. To solve this, ARM algorithms use support and confidence *thresholds*. Only the items with a support value greater or equal than a fixed *minimum support* are used to perform the analysis. Only those rules with confidence value greater or equal than a fixed *minimum confidence* are discovered by the analysis.

ARs have been extensively used during the last two decades. Liu et al. (1998) propose to build classifiers based on the ARM’s *support* and *confidence* metrics and create the first Classification Association Rules Mining (CARM) algorithm, namely the Classification Based on Associations (CBA). A proposal by Li, Han, and Pei (2001) outperforms CBA algorithm using more CARs to build the classifier. Yin and Han (2003) realise that previous CARM algorithms generate a very large number of association rules. Coenen and Leng (2004, 2007) write an in-depth survey on CARM methodologies and propose the use of LUCS-KDD (Coenen, 2003) to discretise numeric attributes in a fixed number of regions. Thabtah, Cowling, and Hamoud (2006) improve rule sorting of previous algorithms and scan only once the dataset in order to improve also time consuming tasks. Hernández-León, Carrasco-Ochoa, Martínez-Trinidad, and Hernández-Palancar (2012) expose how CARs are increasingly applied over different research areas. Pinho Lucas, Segrera, and Moreno (2012) use CARs for a Recommender System. Nguyen, Nguyen, Vo, and Hong (2015) and Nguyen, Nguyen, Vo, and Pedrycz (2016) use class constraints to reduce the amount of data to analyse. Khanmohammadi and Chou (2016) highlight that rule-based classifiers can not directly handle numerical data, and apply a new method for discretization over medical data. Song and Lee (2017) propose a new rank-

[*] Corresponding author. Tel.: +34 966 658 725; fax: +34 966 658 715.
E-mail addresses: enrique@umh.es (E. Lazcorreta Puigmartí), federico@umh.es (F. Botella), Antonio.Fdez@uclm.es (A. Fernández-Caballero).

https://doi.org/10.1016/j.eswa.2017.10.029
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

ing metric for the CARs discovered in a dataset. All of them use three steps proposed by Liu et al. (1998): (1) discretising numeric attributes, if any, (2) generating all the CARs with minimum support threshold, and (3) building a classifier based on ranking generated CARs. The experiments performed in these works support the good results of CARM techniques when compared with the existing ones in ML, concretely with C4.5 (Quinlan, 1993). All the previously cited papers use datasets from the UCI Machine Learning Repository (UCI, 2013) or the KEEL Standard Classification Datasets Repository (KEEL Standard Dataset, 2004) to carry out the comparison.

Despite the small size of the datasets used in the previous works, the minimum support threshold is used to complete CARM algorithms. Coenen and Leng (2007) study the effect that thresholds in support and confidence have on CARM experimentation. Rai, Verma, and Thoke (2012) introduce ARM techniques to discover more knowledge about rare items. All of them use minimum support and, thus, forget some of the information contained in the datasets. Lazcorreta Puigmartí, Botella, and Fernández-Caballero (2012) add information about the classification problem to the dataset and reduce its dimensions, removing the rare item dilemma in some datasets from KEEL.

Support of an item represents the actual frequency of the item in a population. So, the minimum support criterion allows to get a representation of the majority of the population under study. ARM datasets should be representative samples, so that the representativeness of the support metric is accepted. But classification datasets are either representative samples or catalogues, that is, lists containing different records from a sample. Usually, a classification experiment starts taking a representative sample from a population and storing all the gathered evidences in a dataset. But, in some cases, the researchers are only interested in creating a catalogue, so they do not store duplicates in the dataset. Both datasets contain a lot of information on the classification problem, but catalogues have no knowledge about frequency distributions of the items in the population, and the support of a rule must not be interpreted as usual.

When datasets from UCI or KEEL repositories are used, researchers do not know if they are working with samples or with catalogues. Even if they know that a dataset is a sample, do not know if the sample is representative in order to use the support as a good estimator of the population frequency. In any case, removing duplicates from the dataset, a catalogue is obtained. Researchers will get best knowledge about the original dataset using our proposal to analyse catalogues.

In this paper, the *Analysis of Catalogue Datasets based on Characterisations* ($\mathcal{ACDC}$) algorithm is introduced. The algorithm is tested with the 75 datasets from KEEL Standard Dataset (2004) without using minimum support, i.e., using every data stored in each dataset. The complete set of values measured in each individual sample is its *characterisation*, and the pair formed by a characterisation and the class of an individual is a gathered *evidence*. This algorithm does not separate values in the way ARM methods do. So, the number of different data analysed by the algorithm is reduced to the number of individual samples in the dataset. The $\mathcal{ACDC}$ algorithm discovers complete sets of evidences without uncertainty in the dataset, characterisations labelled with an unique class that are named *robust evidences*. Each robust evidence is actually an association rule extracted from the dataset with 100% confidence level. The experiments performed in this paper demonstrate that datasets contain millions or even billions of rules with 100% confidence. The algorithm does not look for all the robust evidences in the dataset, but just looks for robust catalogues, sets of robust evidences based on a unique subset of the original set of attributes.

The rest of the paper is organised as follows. In the Section 2, the theoretical model used for the definition of the proposed algorithm is presented. Section 3 describes the proposed $\mathcal{ACDC}$ algorithm. In Section 4 the results of our experiments are shown. Finally, Section 5 outlines the conclusions and future work.

2. Theoretical basis of the proposed algorithm

The $\mathcal{ACDC}$ algorithm is based on set theory. Datasets are sets composed by data stored with a fixed structure. Classification datasets store data about the measurement of N attributes in an individual, and data about the classification of this individual. These datasets are really matrices with $N+1$ columns and a row for each individual in the sample or catalogue. In this section, the theoretical model used for the definition of the $\mathcal{ACDC}$ algorithm is exposed.

Let $\mathcal{C}=\{c_1 \ldots c_Q\}$ be a set of *classes* partitioning the population under study, where each individual belongs to one and only one class. Let $\mathcal{A}=\{A_1 \ldots A_N\}$ be an ordered set of measurable *attributes*, and *characterisation* of an individual the vector $x=(A_1=x_1 \cdots, A_N=x_N)=(x_1,\ldots,x_N)$, where x_i is a value in the range of A_i. A characterisation from already qualified individuals is an *evidence*. A set of evidences is called *standard classification dataset* (dataset from now on). The classification problem aims to learn the best strategy from the data stored in the datasets in order to classify any new individual from this population by using only its characterisation.

As discussed in Section 1, it is usually necessary to reduce the size of a dataset in order to discover the information contained. The CARM approach suggests to dispense the items with less support in the dataset. Our approach is based on a better understanding of the dataset to be explored before removing a single data. By applying set theory and generic characteristics to all datasets, redundant information is discovered.

Before analysing a dataset in an unsupervised manner, it has to be understood that it could have different characteristics. It can contain, or not, (1) uncertainty, (2) evidences with unknown data (incomplete evidences), (3) duplicate evidences, and (4) non-variable attributes. None of them are considered in ARM algorithms. Our $\mathcal{ACDC}$ algorithm incorporates all of them in the analysis; it is a pre-processing of the dataset that provides other smaller datasets with the same information for classification.

The $\mathcal{ACDC}$ proposed algorithm analyses datasets without incomplete or duplicate evidences, i.e. catalogues, and takes into account the presence of non-variable attributes.

Definition 1 (Catalogue)**.** A *Catalogue* $\mathcal{D}$ is a classification dataset with no incomplete or duplicate evidences.

A catalogue does not contain the same information as a classification dataset. However, its analysis provides valuable knowledge about the classification problem. Any dataset with incomplete and/or duplicate evidences contains a catalogue $\mathcal{D}$that is analysable by our algorithm.

$$\begin{aligned} dataset &= \mathcal{D} \cup \{incomplete\text{-}evidence\} \\ &\quad \cup \{duplicate\text{-}evidence\text{-}removed\} \end{aligned} \tag{1}$$

Once duplicates have been removed from the dataset, the *catalogue* obtained has no knowledge about the original support of each item in the dataset, and must not use this (unknown) support as an estimator of the frequency of any item in the population, as CARM algorithms use to do (Balaji & Rao, 2013; Coenen & Leng, 2007; Kundu, Munir, Bari, Islam, & Murase, 2008; Liu et al., 1998; Song & Lee, 2017).

Lemma 1 (Meaning of the support metric in catalogues)**.** *A Catalogue $\mathcal{D}$does not contain information about the frequency distribution of the population from which it has been obtained. Thus,* support *is*