Expert Systems With Applications 109 (2018) 114–130

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Study of data transformation techniques for adapting single-label prototype selection algorithms to multi-label learning

Álvar Arnaiz-González, José-Francisco Díez-Pastor, Juan J. Rodríguez, César García-Osorio*

Escuela Politécnica Superior, Universidad de Burgos, Burgos 09006, Spain

ARTICLE INFO

Article history:
Received 5 September 2017
Revised 26 April 2018
Accepted 15 May 2018

Keywords:
Multi-label classification
Prototype selection
Binary relevance
Label powerset
RA*k*EL

ABSTRACT

In this paper, the focus is on the application of prototype selection to multi-label data sets as a preliminary stage in the learning process. There are two general strategies when designing Machine Learning algorithms that are capable of dealing with multi-label problems: data transformation and method adaptation. These strategies have been successfully applied in obtaining classifiers and regressors for multi-label learning. Here we investigate the feasibility of data transformation in obtaining prototype selection algorithms for multi-label data sets from three prototype selection algorithms for single-label. The data transformation methods used were: binary relevance, dependent binary relevance, label powerset, and random *k*-labelsets. The general conclusion is that the methods of prototype selection obtained using data transformation are not better than those obtained through method adaptation. Moreover, prototype selection algorithms designed for multi-label do not do an entirely satisfactory job, because, although they reduce the size of the data set, without affecting significantly the accuracy, the classifier trained with the reduced data set does not improve the accuracy of the classifier when it is trained with the whole data set.

1. Introduction

Classification processes are present in a wide range of human activities. The classification task is easy to understand: make a decision or a prediction for a new example or situation on the basis of available information. Focusing on Machine Learning classification, the aim is to train a model from a set of examples, in such a way that the model is capable of making predictions for a new and hitherto unseen example.

Normally, a single label is assigned to each instance, which is known as single-label classification. If there are only two possible labels, it becomes a binary problem; if there are more, then it is a multi-class problem. Moreover, multi-label classification has appeared as a more recent term in the past few decades, in reference to a situation in which more than one label can be assigned to an instance (Zhang & Zhou, 2014). We can think of multi-label classification as a process in which labelsets are assigned that contain several labels rather than single labels. For example, the labels in an image classification problem, such as sky, plane, and cloud, can be assigned to the same instance. This property makes the prediction process more challenging, due to the existence of several labels for each instance (Younes, Abdallah, Denœux, & Snoussi, 2011). Despite the fact that multi-label emerged mainly for medical diagnosis and text categorization, where a document can belong to a set of different topics (Tsoumakas & Katakis, 2007), there are many other applications where its utility is recognized. The following section gathers together both algorithms and multi-label learning applications.

Pre-processing techniques, in both multi-label and conventional single-label, are essential in the Machine Learning workflow. According to García, Luengo, and Herrera (2014), pre-processing tasks are commonly divided into: data preparation, and data reduction. Data preparation comprises data normalization, data cleaning and noise detection, among others; while data reduction, as the name implies, reduces the overall volume of data, while preserving the essential information. These methods can be categorized into three groups: discretization, feature selection, and instance selection (Herrera, Charte, Rivera, & del Jesus, 2016). Commonly, multi-label data sets are high dimensional data sets, which is why feature selection has been widely researched for multi-label data sets (Lee & Kim, 2015; Pereira, Plastino, Zadrozny, & Merschmann, 2016; Spolaôr, Monard, Tsoumakas, & Lee, 2016).

* Corresponding author.
E-mail addresses: alvarag@ubu.es (Á. Arnaiz-González), jfdpastor@ubu.es (J.-F. Díez-Pastor), jjrodriguez@ubu.es (J.J. Rodríguez), cgosorio@ubu.es (C. García-Osorio).

https://doi.org/10.1016/j.eswa.2018.05.017
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

Conversely, instance selection[1] is an underexplored research area in multi-label learning. This aspect of data dimensionality presents a problem, because instance selection methods are also important and useful for reducing and cleaning multi-label data sets (Kargar-Shooroki, Chahooki, & Javanmardi, 2015).

The main contributions of this paper can be summarized as follows:

- For the first time several data transformation methods are used to obtain new algorithms for prototype selection for multi-label data sets from the corresponding prototype selection algorithms for single-label learning. The data transformation techniques used are: binary relevance (Godbole & Sarawagi, 2004), dependent binary relevance (Montañes et al., 2014), label powerset (Boutell, Luo, Shen, & Brown, 2004), and random *k*-labelsets (RA*k*EL) (Tsoumakas, Katakis, & Vlahavas, 2011). The single-label prototype selection algorithms are: Wilson Editing (ENN) (Wilson, 1972), RNGE (Sánchez, Pla, & Ferri, 1997), and local set-based smoother (LSSm) (Leyva, González, & Pérez, 2015).
- An experimental study is carried out to evaluate the performance of the 12 new prototype selection algorithms. They are compared among them, with two algorithms of instance selection obtained using method adaptation, and also with the results of a classifier trained on the original data set.

The paper is structured as follows: Section 2 briefly describes multi-label learning algorithms and their applications; Section 3 presents some background to prototype selection, with special emphasis on the lack of methods for multi-label; Section 4 describes the proposed meta-models based on data transformation; the experimentation details and the results are shown in Section 5. Finally, Section 6 presents the main conclusions of this paper and Section 7 outlines future lines of research.

2. Multi-label learning

As stated before, multi-label learning uses instances/prototypes to which more than one label may have been assigned, or equivalently, instead of being assigned just one label, they are assigned what is called a labelset. The goal of multi-label classification is to construct a predictive model that will be able to produce a set of labels for a given example never seen before (Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012). The simultaneous presence of several labels to the instances of multi-label data sets is what makes their use in learning tasks more challenging than when single-label sets are used (Elisseeff & Weston, 2001; Zhang & Zhou, 2007).

2.1. Multi-label algorithms

Two main approaches have been applied to deal with multi-label data sets: data transformation and method adaptation. Data transformation consists of replacing the multi-label data set by one or more single-label data sets, obtained from the transformation of the original data set. The idea of method adaptation is to modify single label methods, so that they can be used directly in data sets with multiple labels. Some algorithms, initially designed for binary classification, have previously been adapted to multi-class data sets in the same way. Examples of method adaptation are, among others, binarization, voting methods, and divide-and-conquer procedures (Herrera et al., 2016). Due to restrictions on the length of this paper, it is not possible to go into the detail of all the methods proposed in each of these two approaches. We would recommend the works of Zhang and Zhou (2014), Tsoumakas and Katakis (2007), and Herrera et al. (2016) for those readers with an interest in the matter.

Two common simple strategies are used for data transformation: label powerset, and binary relevance. The former considers each labelset (set of labels) as a class in itself, so the initial multi-label data set is transformed into a single-label one with as many classes as the number of labelsets present in the original set (Tsoumakas, Katakis et al., 2011). Fig. 1 shows a tabular representation of a multi-label data set on the left, and the data set generated by label powerset transformation on the right. The drawback of the above approach is that the resultant data set can easily become imbalanced, because some combinations may be poorly represented and the total number of different combinations increases exponentially.

The other strategy, and the most intuitive approach to the data of transformation, is to decompose the data set into several independent single-label data sets, one for each label (Tsoumakas & Katakis, 2007). The binary relevance decomposes a multi-label data set with *n* different labels into *n* single-label data sets. Each new data set contains information of only one label and ignores the rest of them. All new data sets have the same number of instances as the original data set, but with only one label, i.e., the resultant data sets are no longer multi-label but single-label. This is an idea similar to the one used for single-label to transform a multi-class problem into several binary classification problems (one-vs-all), when the method we want to apply is not able to deal with more than one value for the class. This naïve idea is called *binary relevance*, and it is the principal baseline for multi-label classification (Luaces, Díez, Barranquero, del Coz, & Bahamonde, 2012). Its main advantages are its low computational complexity, its simplicity (it is possible to use any single-label learner) and, despite criticism over the assumption of label independence, its effectiveness at producing good multi-label classifiers as it has been shown in several papers (Luaces et al., 2012; Sucar et al., 2014). Fig. 2 shows the results of using binary relevance on the table shown in Fig. 1(a).

A common drawback of the binary relevance technique is that it ignores the relation between labels, which is usually of some importance in multi-label learning. For this reason, recent research on this method has produced an interesting extension called *dependent binary relevance* (Montañes et al., 2014), which employs an extended set of features with additional information on the labelsets. Fig. 3 graphically shows how this works on the data set of Fig. 1(a). As the original binary relevance transformation does, it generates as many single-label data sets as labels present on the original multi-label data set. In contrast to the common binary relevance method, the rest of labels are not ignored, but used as input features.

As it has been explained, label powerset transforms a multi-label data set into a multi-class single-label data set (one class for each labelset). This approach serves as the foundation of many multi-label classifiers ensembles. One of the most used is random *k*-labelsets, or RA*k*EL for short (Tsoumakas, Katakis et al., 2011). RA*k*EL generates m data sets by selecting labelsets with k labels (where $k \leq |\Omega|$, being $|\Omega|$ the number of labels of the data set) in a random way. It can be seen as a technique in between binary relevance and label powerset (if $k = 1$ and $m = \Omega$, it behaves as the former; whereas if $m = 1$ and $k = |\Omega|$, the behavior is as the latter). Thus, the most interesting results are achieved with intermediate values.

All of these techniques have been broadly used for the adaptation of single-label learning methods to multi-label data sets. However, to the best of our knowledge, they have never before been used for adapting to multi-label learning instance selection methods.

[1] Both, instance selection and prototype selection, are commonly used for naming this kind of subsampling methods (García, Derrac, Cano, & Herrera, 2012).