Information Sciences 330 (2016) 370–384

Contents lists available at ScienceDirect

Information Sciences

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

LAIM discretization for multi-label data

Alberto Cano[a,b], José María Luna[b], Eva L. Gibaja[b], Sebastián Ventura[b,c,*]

[a] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
[b] Department of Computer Science and Numerical Analysis, University of Cordoba, Spain
[c] Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 6 August 2015
Accepted 18 October 2015
Available online 22 October 2015

Keywords:
Multi-label learning
Data discretization

ABSTRACT

Multi-label learning is a challenging task in data mining which has attracted growing attention in recent years. Despite the fact that many multi-label datasets have continuous features, general algorithms developed specially to transform multi-label datasets with continuous attributes' values into a finite number of intervals have not been proposed to date. Many classification algorithms require discrete values as the input and studies have shown that supervised discretization may improve classification performance. This paper presents a Label-Attribute Interdependence Maximization (LAIM) discretization method for multi-label data. LAIM is inspired in the discretization heuristic of CAIM for single-label classification. The maximization of the label-attribute interdependence is expected to improve labels prediction in data separated through disjoint intervals. The main aim of this paper is to present a discretization method specifically designed to deal with multi-label data and to analyze whether this can improve the performance of multi-label learning methods. To this end, the experimental analysis evaluates the performance of 12 multi-label learning algorithms (transformation, adaptation, and ensemble-based) on a series of 16 multi-label datasets with and without supervised and unsupervised discretization, showing that LAIM discretization improves the performance for many algorithms and measures.

1. Introduction

The machine learning community has studied the classification task in depth. The usual way to define this task involves associating one class label with each pattern. It can be distinguished between binary and multiclass classification. In the former only two classes are defined, indicating if the pattern belongs or not to the target whilst in the latter more than two classes are defined. This classical definition of the task entails the *only-one-label-per-pattern* restriction. Nevertheless there are more and more current classification problems, such as text and sound categorization, semantic scene classification or gene and protein function classification, in which a pattern could have simultaneously associated not one but a set of labels. These problems with multiple outputs entail specific difficulties such as the exponential growth of combinations of labels to take into account, label correlations and even data imbalance. All of these factors have led to the emerging of *Multi-Label Learning* (MLL) paradigm [20,59]. In contrast to classical (a.k.a. single-label) learning, MLL is able to address problems where class labels are not mutually exclusive. First applications of MLL were related to classification of text and multimedia [44,45], in which one document or picture could be simultaneously associated with several categories, and protein and gene function classification, in which a gene

* Corresponding author. Tel.: +34 957212218; fax: +34 957218630.
E-mail addresses: acano@vcu.edu (A. Cano), jmluna@uco.es (J.M. Luna), egibaja@uco.es (E.L. Gibaja), sventura@uco.es (S. Ventura).

http://dx.doi.org/10.1016/j.ins.2015.10.032
0020-0255/

or protein can perform several functions [58]. Nowadays MLL has become a challenging research area with an increasing number of papers and domains of application such as drug discovery [23], social network mining [26], and direct marketing [60].

On the other hand, many machine learning and statistical techniques have been designed to learn only in datasets composed of nominal variables while real-world applications usually involve continuous features [25,28]. In order to overcome this drawback one solution is to use an embedded or external method to discretize continuous features by partitioning them into a number of discrete intervals and treat each one as a category. As it maps from a high dimensional range of values to a reduced subset of discrete values, discretizatcion can be considered a data reduction method [19]. Obtaining the optimal discretization is NP-complete [9] being a potential time-consuming bottle-neck.

Despite the fact that many multi-label datasets have continuous features, general algorithms developed specially to transform multi-label datasets with continuous attributes' values into a finite number of intervals have not been proposed to date. Studies have shown the advantages of supervised discretization [19] and many classifiers require discrete input. This paper presents a discretization approach based on Label-Attribute Interdependence Maximization (LAIM) that can be applied to numerical multi-label datasets. The proposal is inspired on the CAIM [27] discretization method for multi-class single-label classification and extends its application to multi-label data. The label-attribute interdependence maximization is expected to improve the generation of discrete intervals that boost the performance of subsequent multi-label classifiers. The primary objective of this paper is to present a supervised discretization method specifically designed for multi-label data and to analyze whether multi-label discretization can improve performance of state-of-art MLL algorithms. The experimental study evaluates and compares the performance of 12 multi-label algorithms (transformation, adaptation, and ensemble-based) with and without discretization on 16 datasets. Moreover, 13 different metrics are used to evaluate the performance of the algorithms. The experimental results are contrasted through the analysis of non-parametric statistical tests [18], namely the Wilcoxon [53] test that evaluates whether there are statistically significant differences between the performance of algorithms on discretized and non-discretized data.

The paper is structured as follows. Section 2 reviews related works on multi-label and data discretization. Section 3 describes the LAIM discretization method to multi-label data. Section 4 presents the experimental study and Section 5 discusses the results and the statistical analysis. Finally, Section 6 shows the main conclusions of this work.

2. Background

In order to provide the reader with the necessary background, this section presents an overview of both MLL and discretization techniques.

2.1. Multi-label learning

Given $\mathcal{F} = F_1 \times \cdots \times F_k$ a k-dimensional input space of numerical or categorical features, and an output space of q labels, $\mathcal{Y} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$, in MLL an instance has the form $(\mathbf{x}, Y)$, where $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{F}$, and $Y \subseteq \mathcal{Y}$ is called *labelset*. Label associations can be also represented as a q dimensional binary vector $\mathbf{y} = (y_1, y_2, ..., y_q) = \{0, 1\}^q$ where each element is 1 if the label is relevant and 0 otherwise. Note that in single-label learning each example has the form $(\mathbf{x}, y)$, where $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{F}$, and $y \in \mathcal{Y}$.

MLL problems can be dealt from two points of view [47]. On the one hand some studies have proposed *transformation methods*, which transform an original multi-label problem into one or several single-label problems, which will then be resolved using a classical classification algorithm. On the other hand, there are studies proposing the extension of classical classification paradigms to cope with multi-label data directly; these are called *algorithm adaptation methods*.

Some transformation methods are based on label combinations. Thus, the *Label Powerset* (LP) [47] method considers each combination of labels in the original dataset as a new and different label. The main drawback of LP is its complexity, that grows exponentially with the number of labels. The *Pruned Sets* (PS) [39] method is similar to LP but specifically designed for problems with a large number of label combinations. Therefore, PS prunes the patterns associated with the less frequent combinations and after that, it reintroduces the pruned examples along with frequent subsets of their label sets. *Ensemble of Pruned Sets* (EPS) [41] constructs a number of PS by sampling the training sets (i.e. boostrap). The RA*k*EL algorithm [48] produces several LP classifiers, which are specialized in classifying random subsets of labels. Answers are combined by a voting process.

Other transformation methods are based on binary decompositions of the problem. *Binary Relevance* (BR) [47] generates one independent binary classifier for each label, the positive patterns being the ones belonging to the label, and being the rest negative patterns. The final output of the multi-label classifier is obtained by combining the outcomes of all of the classifiers. The main problem of BR is the independence assumption. *Classifier Chains* (CC) [42] also generates q binary classifiers, but they are linked in such a way that the feature space of each link in the chain is extended with the labels associations of all previous links. Thus, CC overcomes the label independence assumption of BR and the potentially computational complexity of LP. *Ensemble of Classifier Chains* (ECC) trains a set of CC classifiers with a random chain ordering and a random subset of training patterns. *Dependent Binary Relevance* (DBR) [34] also follows a binary decomposition but incorporating, for each binary classifier, the information of the rest of labels as additional features.

Calibrated Label Ranking (CLR) [5] carries out a pairwise decomposition and produces a binary model for each pair of labels. Besides, it adds a virtual label that is used as a splitting point between positive and negative labels.

Regarding the problem transformation approach, many single-label algorithms have been adapted to deal with multi-label data directly, without pre-processing the multi-label dataset. For instance, decision trees [10], SVMs [33,51], associative