Neurocomputing 196 (2016) 113–124

Contents lists available at ScienceDirect

Neurocomputing

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

Positive and unlabeled learning in categorical data

Dino Ienco [a,c,1], Ruggero G. Pensa [b,*]

[a] IRSTEA Montpellier, UMR TETIS, F-34093 Montpellier, France
[b] Department of Computer Science, University of Torino, I-10149 Torino, Italy
[c] LIRMM Montpellier, ADVANSE, F-34090 Montpellier, France

ARTICLE INFO

Article history:
Received 26 January 2015
Received in revised form
23 October 2015
Accepted 10 January 2016
Available online 9 March 2016

Keywords:
Positive unlabeled learning
Partially supervised learning
Distance learning
Categorical data

ABSTRACT

In common binary classification scenarios, the presence of both positive and negative examples in training data is needed to build an efficient classifier. Unfortunately, in many domains, this requirement is not satisfied and only one class of examples is available. To cope with this setting, classification algorithms have been introduced that learn from Positive and Unlabeled (PU) data. Originally, these approaches were exploited in the context of document classification. Only few works address the PU problem for categorical datasets. Nevertheless, the available algorithms are mainly based on Naive Bayes classifiers. In this work we present a new distance based PU learning approach for categorical data: *Pulce*. Our framework takes advantage of the intrinsic relationships between attribute values and exceeds the independence assumption made by Naive Bayes. *Pulce*, in fact, leverages on the statistical properties of the data to learn a distance metric employed during the classification task. We extensively validate our approach over real world datasets and demonstrate that our strategy obtains statistically significant improvements w.r.t. state-of-the-art competitors.

1. Introduction

In common binary classification tasks, learning algorithms assume the presence of both positive and negative examples. Sometimes this is a strong requirement that does not fit real application scenarios. In fact, the process of labeling data is a money- and time-consuming activity that needs high-level domain expertise. In some cases this operation is quick, but usually, defining reliable labels for each data example is a hard task. In the worst case, extracting examples from one or more classes is simply impossible [1]. As a consequence, only a small portion of a so-constituted training set is labeled. As a practical example of this phenomenon, let us consider a company that aims at creating an archive of researchers' home pages, using web-crawling techniques. Once downloaded, a web page should be classified to decide whether it is a researcher's home page or not. In such a context, the concept of positive example is well defined (the researcher's home page) while the idea of negative example is not well-established [2] because no real characterization of what is not a home page is supplied. The same problem occurs when trying to classify biological/medical data. Usually a biologist (or a doctor) can comfortably supply positive evidences of what she wants to identify but she is not able to provide negative examples. A known example of this scenario is the classification of vascular lesions starting from medical images [3], where labeling vascular lesions accurately could take more than one year, while it is relatively easy to recognize healthy individuals. In these scenarios, defining a method to exploit both positive and unlabeled examples could save precious material and human resources and the expert may focus her effort to only define what is good, skipping the ungrateful task of recognizing what is not good.

To deal with this setting, the Positive Unlabeled (PU) learning task has been introduced [4]. Roughly speaking, PU learning is a binary classification task where no negative examples are available. Most research works in this area are devoted to the classification of unstructured datasets such as documents represented by bag-of-words, but similar scenarios may occur with categorical data as well. Imagine, for instance, a dataset representing census records on a population. An analyst can comfortably provide reliable positive examples of a targeted class of people (e.g., unmarried young professional interested in adventure sports), but identifying plausible counterexamples is not as easy. However, very few PU learning approaches are designed to work on attribute-relation data (such as categorical datasets). Unfortunately the techniques proposed in text classification are not directly applicable to the context of attribute-relation datasets. These approaches, in fact, employ metrics, such as the cosine distance, that are not well suited for categorical data where, in addition, there is no standard definition of distance [5]. This limitation makes it impossible to apply works on document classification to categorical data directly. The few works that deal with PU learning in attribute-relation domains are principally based

* Corresponding author. Tel. +39 011 670 6798; fax: +39 011 75 16 03.
E-mail addresses: dino.ienco@irstea.fr (D. Ienco), pensa@di.unito.it (R.G. Pensa).
[1] Tel. +33 4 67 55 86 12; fax: +33 4 67 54 87 00.

http://dx.doi.org/10.1016/j.neucom.2016.01.089
0925-2312/

on Naive Bayes classifiers. The major limitation of this kind of approaches is that algorithms based on Naive Bayes assume that attributes are mutually independent. To the best of our knowledge, no effort was devoted to the implementation of other models or the extension of previously defined models from document analysis.

In this paper we introduce a new distance-based algorithm, named *Positive Unlabeled Learning for Categorical datasEt* (*Pulce*). Our work aims at filling the gap between the recent and well-established advances in document classification and the preliminary status of works existing for attribute-relation data. In particular, we address the problem of classifying data described by categorical attributes, which also includes the case of discretized numerical attributes, leading to a general framework for attribute-relation data. The core part of our approach is an original distance-based classification method which employs a distance metric learnt directly from data thanks to a technique recently presented by Ienco et al. [6]. Originally, this technique was designed to exploit attribute dependencies in an unsupervised (clustering) scenario. Its key intuition is that the distance between two values of a categorical attribute X_i can be determined by the way in which they co-occur with the values of other attributes in the dataset: if two values of X_i are similarly distributed w.r.t. other attributes X_j (with $i \neq j$), the distance is low. The added value of this proximity definition is that it takes into consideration the *context* of the categorical attribute, defined as the set of the other attributes that are relevant and nonredundant for the determination of the categorical values. Relevancy and redundancy are determined by the symmetric uncertainty measure that is shown to be a good estimate of the correlation between attributes [7].

Our PU learning approach uses this metric to train two discriminative models: one for the positive class, the other for the negative one. These two models take intrinsically into account the existing attribute relationships, thus overcoming the major limitation given by the independence assumption explicitly made by Naive Bayes-based methods. We provide the empirical evidence of this property, showing that our method outperforms state-of-the-art competitors and assessing the statistical significance of the results. In a nutshell, our contributions can be summarized as follows:

- we introduce a distance learning approach to detect reliable negative examples in datasets described by categorical attributes;
- we leverage the same distance learning approach to build two distance models: one for the positive examples, one for the negative ones;
- we define a *k*-NN classifier to predict the positive/negative label of the unseen examples: each example is assigned the class label of the distance model (positive/negative) it fits better.
- we compare our approach to other recent state-of-the-art PU classification methods and show a statistically significant improvement in terms of prediction rates.

The remainder of this paper is organized as follows: Section 2 briefly explores the state-of-the-art in PU learning and other close research areas. The problem formulation, a brief overview of the distance learning algorithm, and the full description of the proposed method are supplied in Section 3. In Section 4 we provide our empirical study and analyze its statistical significance. Finally, Section 5 concludes.

2. Related work

Positive Unlabeled learning was originally studied by De Comité et al. [4] who achieved the first theoretical results. The authors showed that under the PAC (Probably Approximately Correct) learning model, the k-DNF (k-Disjunction Normal Form) approach is able to learn from positive and unlabeled examples. Following these preliminaries results, PU learning was first applied to text document classification [2]. In this work the authors design a method that uses 1-DNF rules to extract a set of reliable negative examples. Then, they use an approach based on support vector machines to learn a classification model over the set of positive and reliable negative examples. The proposed technique achieves the same performances as classical SVMs do for the web page classification task.

Other approaches dealing with PU classifiers in the context of text classification have been presented in more recent years [8–10]. Elkan et al. [8] introduce a method to assign weights to the examples belonging to the unlabeled set. The whole set of weighted unlabeled examples is then used to build the final SVM-based classifier. Also Xiao et al. [9] present an approach based on SVMs. The authors combine two techniques borrowed from information retrieval (Rocchio and Spy-EM) to extract a set of reliable negative examples. Then a weighting schema is applied on the remaining unlabeled examples. To exploit these three sources of information (positive, reliable negative and weighted unknown examples) the authors adapt standard SVMs. Finally, Zhou et al. [10] tackle the problem from a different point of view: they modify the standard Topic-Sensitive probabilistic Latent Semantic Analysis (pLSA) approach to perform classification with a small set of positive labeled examples. The information carried by the positive class is used to constrain the unsupervised process usually adopted by pLSA. Because of that, the developed method is more similar to constrained clustering techniques rather than standard PU learning. In fact, no real learning algorithm is involved and the final result is not a classifier.

Recently, the scientific literature concerning PU learning systems has been enriched of some new theoretical results [11,12] and new successful applications [13,14]. Mordelet et al. [11] propose a new method for PU learning based on bagging techniques, while Wu et al. [12] propose a SVM-based solution to the problem of positive and unlabeled multi-instance learning. Li et al. [13] propose a collective classification approach from positive and unlabeled examples to identify fake reviews. In a completely different domain, Yang et al. [14] use a similar approach to identify causative genes to various human diseases.

Differently from document classification, the literature on PU learning for categorical data is not as rich. This is due to the fact that no consensus on how to evaluate distances in categorical data has been reached yet. In fact, while in document classification standard objective measures as cosine or Euclidean distance are widely employed, this is not the case for categorical data, where distance measures are mainly based on extracted statistics depending on the specific dataset [5].

Calvo et al. [15] first attempted to deal with the PU learning setting in attribute-relation datasets. Their paper introduces four methods based on Naive Bayes for categorical data. In particular the authors modify classic and Tree Augmented Naive Bayes [16] approaches to work with positive and unlabeled examples. They supply two ways to estimate the prior probability of the negative class: the first one takes into consideration the whole set of unlabeled examples to derive this probability, while the second one considers a Beta distribution to model the uncertainty. These methods are substantially limited by two aspects: the strong (and often wrong) assumption of attribute independency adopted by Naive Bayes and the use of the whole set of unlabeled examples to estimate a model for the negative class. This work is extended by He et al. [17] to deal with uncertainty data. Another attempt aiming at bringing the PU learning problem outside text document classification has been presented by Zhao et al. [18]. Their work provides a formalization of PU learning for classifying graphs via an optimization strategy that tries to learn a good classifier and a good set of discriminant graph features from both positive and unlabeled examples. Though effective on graphs, unfortunately, this method cannot be adapted to other data types. Very recently, Shao et al. [19] proposed a novel PU learning approach called Laplacian Unit-Hyperplane