Pattern Recognition 46 (2013) 3425–3440

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Learning Bayesian network classifiers from label proportions

Jerónimo Hernández-González *, Iñaki Inza, Jose A. Lozano

Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU
Paseo Manuel de Lardizabal 1, 20018 Donostia-San Sebastián, Spain

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form
25 April 2013
Accepted 2 May 2013
Available online 15 May 2013

Keywords:
Supervised classification
Learning from label proportions
Structural EM algorithm
Bayesian network classifiers

ABSTRACT

This paper deals with a classification problem known as learning from label proportions. The provided dataset is composed of unlabeled instances and is divided into disjoint groups. General class information is given within the groups: the proportion of instances of the group that belong to each class.

We have developed a method based on the Structural EM strategy that learns Bayesian network classifiers to deal with the exposed problem. Four versions of our proposal are evaluated on synthetic data, and compared with state-of-the-art approaches on real datasets from public repositories. The results obtained show a competitive behavior for the proposed algorithm.

1. Introduction

In classical supervised classification, the objective is to build a predictive model from a dataset of labeled instances such that, given a new unlabeled example, the model will assign it to one of the already-known class labels. In the most common situation, each instance in the dataset consists of the description of the example and its associated class label [1]. Moreover, other problems where obtaining labeled examples is difficult (semi-supervision) have received considerable attention in the literature [2]. However, in recent years new problems in which the available class-membership information of the provided examples (a.k.a. information of supervision) does not consist of the typical class value for each (labeled) instance have been proposed. Thus, standard learning strategies, which have been developed for learning from supervised or semi-supervised domains, cannot be straightforwardly applied. Therefore, new specific strategies (or adaptations of classical strategies) that learn from the new kinds of non-fully labeled datasets are necessary. The specific techniques, in order to be efficient, are expected to extract as much knowledge as possible from the available information of supervision.

In this paper, we deal with a problem in which the relation between an instance and its associated class label is lost. This may be due to the black-box nature of the problem, privacy preserving, non-monitoring process, etc. In this framework, the unlabeled instances are grouped and only global class information is available for the instances of each group: the label proportions. A particular application of this general framework is the problem of embryo selection in *assisted reproductive technology* (ART) [3]. In the most critical step of an ART cycle, gynecologists have to select the embryos to be transferred to the uterus of the woman among a set of embryos that have been cultured for several days (in Spain, by law, three embryos at most can be transferred in an ART cycle). During the culture period, some relevant features are observed and collected for each individual embryo. Then, after the transference, doctors can observe, using preclinical imaging techniques, the number of those transferred embryos that are implanted (and induce a pregnancy), but it is not possible to know *which* individual embryo is implanted. Thus, in a dataset for this problem, each instance represents a transferred embryo and each group includes the embryos transferred in the ART cycle that it represents. The class label, which should indicate whether or not the specific transferred embryo became implanted or not, is individually unknown for the instances of the dataset. However, some kind of information of supervision is available for each group of instances: the number of positive instances (implanted embryos) in the corresponding ART cycle.

Another real case that involves the same kind of data is that of election votes, where some parties stand for institutions and, in each polling station, each party gets a known number of votes. The global election results are known, but which party each citizen voted for is unknown. By knowing the population census and some socioeconomic data of the voters, it could be possible to estimate the probability of a citizen voting for a party [4]. More real instances of the problem include the analysis of single particle mass spectrometry data [5], e-commerce [6], spam and image filtering [6], fraud detection [7], etc.

* Corresponding author. Tel.: +34 943018070.
E-mail addresses: jeronimo.hernandez@ehu.es (J. Hernández-González), inaki.inza@ehu.es (I. Inza), ja.lozano@ehu.es (J.A. Lozano).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.05.002

The presented problem relates to the multiple-instance learning problem since, in both cases, the training dataset is divided into disjoint groups of instances. Multiple-instance learning (MIL) [8] is a supervised classification problem where an example is represented by a group of instances and there is a global label per group (or example). In MIL, the objective is to learn from and classify groups of instances. However, the problem we are dealing with considers class label assignments to the individual instances, despite being unknown in training time.

There exist in the literature several methods to deal with the learning from label proportions (LLP) problem. The first time that a method was proposed to learn from this kind of data was in [4], where Kück and Freitas present a MCMC strategy. But it was Musicant et al. [5] who gave the first definition of the LLP problem, which they called aggregated outputs. They use the counts of labels (instead of proportions) as general class information per group. In their paper, basic adaptations of KNN, ANN, SVM and Decision Trees are proposed.

Simultaneously, Quadrianto et al. [6] gave an alternative definition based on label proportions. Their method, called MeanMap, models the conditional class probability using conditional exponential models. Although their method is primarily defined to deal with problems where the label proportions of the test set are known, it incorporates a functionality that estimates these proportions when they are not given. Following a similar definition of the problem but without requiring label proportions of the test set, Rueping [7] proposes an algorithm to learn SVMs for this problem.

Other authors implement a different strategy to learn from LLP datasets. Their contributions consist of a procedure that firstly reduces the uncertainty of the data provided, estimating the class label of each unlabeled instance. This generates a complete dataset which can be used to train a classifier using any classical method for supervised data. In this way, Chen et al. [9] proposed a method based on kernel K-means for solving this problem of label assignment. Later, Stolpe and Morik [10] presented a similar method which solves this problem using an evolutionary strategy that looks for the predictive variable weights that lead to the clustering (K-means) that best fits the label proportions.

The main contributions of this paper are as follows:

- The development of an algorithm based on the Structural Expectation-Maximization (SEM) strategy [11] to learn Bayesian network classifiers for the LLP problem.
- The development of several variants of the method, two of which have been specifically designed to deal with (complex) LLP scenarios with high degree of uncertainty in the class label of the individual instances.
- The use of joint label assignments, i.e. only the label assignments which fulfill the label proportions of the groups are considered.
- The proposal of a new framework for testing LLP methods, which covers the whole spectrum of LLP scenarios in terms of complexity for a given dataset.

The Bayesian network classifiers that our SEM method learns show a good performance behavior through different LLP scenarios of increasing class uncertainty. Moreover, it obtains competitive results with respect to state-of-the-art methods.

The rest of the paper is organized as follows. In the next section, a formal description of a LLP problem is given. Then, three Bayesian network classifiers and the Structural Expectation-Maximization (SEM) strategy, the basic methodologies implemented in our proposals, are explained. Later, four versions of a new algorithm based on the Structural EM strategy which learns Bayesian network classifiers in the LLP framework are proposed. In Section V, the experiments are presented in four subsections: an experimental demonstration of the usefulness of the extra class information provided in the LLP problems using the semi-supervised learning approach as a baseline-performance reference, an evaluation of the approximate reasoning of our method by means of local probabilistic label assignments, an analysis on synthetic data that evaluates the efficacy of our proposals in different experimental conditions, and a comparison with state-of-the-art approaches. Finally, some conclusions and future work are presented.

2. The problem of learning from label proportions

In a problem of learning from label proportions (LLP), the examples are provided unlabeled and grouped in *bags*—or disjoint sets of examples. In fact, the labels of the instances are known but, for some reason, the individual pairing relation (instance, label) is lost. In this way, a bag is composed of two equal-size unpaired groups: a group of instances and a group of labels. The group of labels can be presented as the proportion of instances that belong to each class label. Note that these label proportions do not indicate a belief (probability) in the number of instances that belong to each class but the real exact number.

A classical supervised learning problem is described by a set of npredictive variables $(X_1, \ldots, X_n)$ and a class variable C. A dataset is a set of examples or instances of the problem, where each instance is a $(n+1)$−tuple that assigns a value to each variable. Specifically, $\mathcal{C}$ represents the set of values, or class labels, that the class variable can take. The objective of supervised classification is to infer the class label of new unlabeled instances.

The dataset D of a LLP problem is composed of m unlabeled examples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. In this paper, we assume that D has been sampled i.i.d. from some underlying probability distribution. The examples are provided grouped in b *bags* $(D = \mathbf{B}_1 \cup \mathbf{B}_2 \cup \cdots \cup \mathbf{B}_b$ where $\mathbf{B}_i \cap \mathbf{B}_j = \emptyset, \forall i \neq j)$. Each bag $\mathbf{B}_i$ groups m_i instances, where $\sum_{i=1}^{b} m_i = m$, and m_{ic} denotes the number of instances in $\mathbf{B}_i$ which have the label c. These m_{ic} values, called *counts* of the bag $\mathbf{B}_i$, sum up to m_i; i.e. $\sum_{c \in \mathcal{C}} m_{ic} = m_i$. Similarly, bag class information can be provided in terms of *proportions* [6], $p_{ic} = m_{ic}/m_i \in [0, 1]$, with $\sum_{c \in \mathcal{C}} p_{ic} = 1$.

Similar to classical supervised classification, the LLP objective remains that of classifying new individual instances, as opposed to multiple-instance learning (MIL), where the objective is to classify new bags (the examples of the problem are groups of instances).

2.1. Uncertainty associated to the label proportions

From the previous description of the LLP problem, the difficulty of pairing each instance with its class label could be thought as a basic definition of uncertainty associated to the label proportions. Thus, assuming that each bag has its own label proportions and, therefore, involves its particular uncertainty, it is possible to distinguish between two kinds of bags. On the one hand, if all the instances in bag $\mathbf{B}_i$ belong to the same class $(\exists c \in \mathcal{C} : m_{ic} = m_i)$, there is class certainty and the individual instances may be considered labeled. This kind of bag is called *full bag*. Following the example of the embryo selection in ART, a bag is full if the corresponding ART cycle finished with either all the embryos implanted or no embryo implanted. However, bags usually have instances that belong to different classes $(\forall c \in \mathcal{C} : m_{ic} < m_i)$. In this case, the class label of an individual instance is unknown (class uncertainty), although the instances in $\mathbf{B}_i$ are known to belong to one of the class labels specified in the label proportions of $\mathbf{B}_i$. This kind of bags are known as *non-full bags*. Following the previous example, this case is observed when some of the transferred embryos (but not all of them) became implanted. It is important to note that the