Information Sciences 436–437 (2018) 131–145

Contents lists available at ScienceDirect

Information Sciences

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

Non-parametric Bayesian annotator combination

M. Servajean [a,b,c,*], R. Chailan [d], A. Joly [b]

[a] Université Paul Valéry Montpellier, Route de Mende, Montpellier 34199, France
[b] Zenith Team from INRIA at LIRMM, 860 rue de St Priest, Montpellier 34095, France
[c] ADVANSE Team at LIRMM, France
[d] Twin Solutions, 11, rue Dulong, Paris 75017, France

ARTICLE INFO

Article history:
Received 6 March 2017
Revised 8 January 2018
Accepted 11 January 2018
Available online 11 January 2018

Keywords:
Bayesian
Variational
Laplace
Classification
Combination
Dirichlet process
Crowdsourcing

ABSTRACT

Relying on a single imperfect human annotator is not recommended in real crowdsourced classification problems. In practice, several annotators' propositions are generally aggregated to obtain a better classification accuracy.

Bayesian approaches, by modeling the relationship between each annotator's output and the possible true labels (classes), have been shown to outperform other simpler models.

Unfortunately, they assume that the total number of true labels is known. This is not the case in lots of realistic scenarios such as open-world classification where the number of possible labels is undetermined and may change over time.

In this paper, we show how to set a non-parametric prior over the possible label set using the Dirichlet process in order to overcome this limitation. We illustrate this prior over the Bayesian annotator combination (BAC) model from the state of the art, resulting in the so-called non-parametric BAC (NPBAC).

We show how to derive its variational equations to evaluate the model and how to assess it when the Dirichlet process has a prior using the Laplace method.

We apply the model to several scenarios related to closed-world classification, open-world classification and novelty detection on a dataset previously published and on two datasets related to plant classification. Our experiments show that NPBAC is able to determine the true number of labels, but also and surprisingly, it largely outperforms the parametric annotator combination by modeling more complex confusions, in particular when few or no training data are available.

1. Introduction

The huge potential of leveraging human power has been noted in recent years, especially when typical machine learning techniques such as deep learning [23] fails. This is particularly needed for instance when constructing the large training sets needed by these techniques [9]. In this paper, we focus on the particular case of data classifications, *i.e.* given a set of data items such as images, sounds or any other documents, we would like to associate a label to each of them. Unfortunately, a perfect annotator that would obtain 100% accuracy does not exist in most realistic scenarios. In practice, we often aggregate information from several annotators, hoping that their abilities are complementary and that the resulting aggregation has better accuracy than a single annotator [23]. Crowdsourcing platforms, such as Amazon Mechanical Turk or Zooniverse,

* Corresponding author at: Université Paul Valéry Montpellier, Route de Mende, 34199 Montpellier, France.
E-mail address: servajean@lirmm.fr (M. Servajean).

https://doi.org/10.1016/j.ins.2018.01.020

offer an efficient way to involve lots of annotators and collect their classification propositions [6,24]. Similarly, in a previous work, we presented The Plant Game, a gamified approach to crowdsourcing where each classification proposition given by an annotator and validated by the crowdsourced consensus increases the annotator's ranking [21]. In all of these platforms, annotators are asked to propose a label without the knowledge of the propositions put forward by other annotators.

A common problem is therefore to merge classification propositions. A simple approach involves counting the number of times each label (class) has been proposed which is called majority voting. More sophisticated methods can be devised such as weighed majority voting [16], where the weight of each annotator depends on its overall classification accuracy. In these approaches, the label of each item stems from the classification propositions. Recent studies have focused on the Bayesian combination of so-called imperfect annotators' propositions [15,17,23]. These models rely on the idea of *confusions* which consists of modeling the output probability of each annotator given all possible true labels – two annotators can have two totally different outputs, such as humans speaking different languages. Therefore they outperform the majority voting approaches. As an example, let us consider a scenario based on ImageNet [9]. The ImageNet challenge consists of finding the true class of a set of images over a large variety of values, such as cars, flowers, etc. An annotator's confusion could indicate that he/she is not capable of disambiguating different types of flowers.

Unfortunately, and contrary to simple majority voting, Bayesian approaches do assume that the number of true labels is known when computing all confusion matrices. This assumption is strong and can be unrealistic in some scenarios. In open-world classification problems [2], determining the set of possible true labels is impossible and can even change over time. In biodiversity surveillance on a crowdsourcing platform, the annotators have to identify species of plants based on their images and they are particularly interested in detecting new species. This would not be possible with a fixed predetermined number of true labels.

In this paper, we propose a non-parametric Bayesian combination model to solve the problem of combining annotators' propositions when the label set is not initially known. In addition, we will show that a non-parametric model enables us to take more complex confusions for each annotator into account.

Related studies focused on Bayesian non-parametric models often rely on a distribution called the Dirichlet process [4,10]. The basic intuition behind such distributions is that the number of possible labels is theoretically infinite while several data items (*e.g.* images, sounds) can have the same label with a positive probability. More formally, the Dirichlet process has infinite dimensions while almost surely staying discrete[1]. The granularity of each resulting class of the non-parametric model depends on the concentration parameter of the Dirichlet process.

However, even though fixing the concentration parameter is less problematic than having to fix the number of possible labels, we also study the model when the concentration parameter itself follows a prior distribution. Thus, the model should converge to the "best" granularity based on the observed data and our prior knowledge.

In order to infer the posterior probabilities, we derive all variational equations required by the model. Variational inference [4,23] is known to approximate the joint probability very efficiently while sampling based methods are known to be much slower [3,23]. Unfortunately, setting a prior over the concentration parameter makes its variational equation intractable. To solve this issue, we show that the Laplace development of the concentration parameter variational equation approximates it by a Gaussian distribution.

In summary, this paper introduces the following original contributions:

- We propose a non-parametric Bayesian annotator combination model to solve the problem of learning the model when the labels set is not known as well as the problem of modeling complex confusions. We also discuss its relationship with the classical parametric model (described in [15,23]).
- We develop variational equations of the non-parametric model in order to efficiently estimate its joint probability, even in high dimensions.
- We show how the Dirichlet process parameter itself can be described with a distribution and how to compute its variational equation using the Laplace method.
- We present an extensive application analysis of previous contributions in the experiments section and show that NPBAC can correctly estimate the number of classes and even outperforms the state of the art Bayesian combination approach that we build upon as well as the simpler majority voting approach.

The rest of this paper is structured as follows. The related work is introduced in Section 2. Section 3 describes the classical parametric Bayesian annotator combination model. In Section 4, we show how to transform the parametric model with the Dirichlet process to make it non-parametric. The variational equations of the non-parametric model are explained in Section 5. In Section 6, we show how to add a prior distribution over the concentration parameter and how to estimate its posterior distribution with the Laplace development of its variational equation. In Section 7 we report and discuss the results of our experiments.

2. Related work

Whereas crowdsourcing is a relatively new domain [6,11,14], contributions related to human classifiers (*i.e.* annotators) combinations or error-rate evaluations go back as far as the 1970s. Dawid and Skene [8], in particular, focused on estimating

[1] "almost surely" refers to the fact that some outcomes, while being theoretically possible, have a zero probability.