Pattern Recognition 42 (2009) 334 – 348

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning from partially supervised data using mixture models and belief functions

E. Côme[a,c,*], L. Oukhellou[a,b], T. Denœux[c], P. Aknin[a]

[a] Institut National de Recherche sur les Transports et leur Sécurité (INRETS) - LTN, 2 av. Malleret-Joinville, 94114 Arcueil Cedex, France
[b] Université Paris XII - CERTES, 61 av. du Général de Gaulle, 94100 Créteil, France
[c] Centre de Recherches de Royallieu, Université de Technologie de Compiègne - HEUDIASYC, B.P. 20529, 60205 Compiègne Cedex, France

ARTICLE INFO

Article history:
Received 13 November 2007
Received in revised form 10 July 2008
Accepted 15 July 2008

Keywords:
Dempster–Shafer theory
Transferable belief model
Mixture models
EM algorithm
Classification
Clustering
Partially supervised learning
Semi-supervised learning

ABSTRACT

This paper addresses classification problems in which the class membership of training data are only partially known. Each learning sample is assumed to consist of a feature vector $\mathbf{x}_i \in \mathcal{X}$ and an imprecise and/or uncertain "soft" label m_i defined as a Dempster–Shafer basic belief assignment over the set of classes. This framework thus generalizes many kinds of learning problems including supervised, unsupervised and semi-supervised learning. Here, it is assumed that the feature vectors are generated from a mixture model. Using the generalized Bayesian theorem, an extension of Bayes' theorem in the belief function framework, we derive a criterion generalizing the likelihood function. A variant of the expectation maximization (EM) algorithm, dedicated to the optimization of this criterion is proposed, allowing us to compute estimates of model parameters. Experimental results demonstrate the ability of this approach to exploit partial information about class labels.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Machine learning classically deals with two different problems: supervised learning (classification) and unsupervised learning (clustering). However, in recent years, new paradigms have emerged to mix these two approaches in order to extend the applicability of machine learning algorithms.

The paradigm that emerged first is *semi-supervised learning* [1,2], where the learning set $\mathbf{X}^{ss} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M), \mathbf{x}_{M+1}, \ldots, \mathbf{x}_N\}$ is composed of two different parts. In the first part, the true class labels y_i are specified, whereas in the second part only the feature vectors $\mathbf{x}_i$ are given. The importance for such problems comes from the fact that labelled data are often difficult to obtain, while unlabelled ones are easily available. Using unlabelled data may thus be a means to enhance the performances of supervised algorithms with low additional cost. The recent publication of a collected volume [3] shows the important activity around this issue in the machine learning field. Recent approaches to semi-supervised learning fall into two main categories:

- An important class of methods is based on the hypothesis that the decision boundary should be located in low density areas. Methods in this category aim at deriving a regularizer of the conditional log-likelihood, taking into account the unlabelled data to bias the decision boundary toward low density areas [4,5]. The transductive support vector machine [6] uses a margin-based criterion to achieve a similar goal. All these methods suffer from the problem of local maxima, although some relaxation schemes lead to a convex optimization problem in the case of the transductive support vector machine [7].
- Other methods are based on the assumption that the high-dimensional input data lie near a low-dimensional manifold. Unlabelled data are then useful as they help in estimating this manifold. Methods relying on the manifold assumption are typically based on unsupervised dimensionality reduction techniques such as PCA or kernel-PCA, or on label propagation in a graph [8,9].

Other paradigms have also been proposed to take into account more sophisticated information on class labels. For example, *partially supervised learning* [10–14] deals with constraints on the possible classes of samples. In this case, the learning set has the following form $\mathbf{X}^{ps} = \{(\mathbf{x}_1, C_1), \ldots, (\mathbf{x}_N, C_N)\}$, where C_i is a set of possible classes for learning example *i*. If all classes are possible, the example is not labelled. Conversely, the example is perfectly labelled if only one class is specified ($|C_i| = 1$). Between these two extreme cases, this approach may also handle situations where some examples are known to belong to any subset of classes. In this case, they are considered as *partially* or *imprecisely* labelled. This framework is thus more general than the semi-supervised learning problem.

[*] Corresponding author at: Institut National de Recherche sur les Transports et leur Sécurité (INRETS) - LTN, 2 av. Malleret-Joinville, 94114 Arcueil Cedex, France. Tel.: +33 1 47 40 73 49.
E-mail address: come@inrets.fr (E. Côme).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.07.014

A completely different paradigm is based on the notion of *label noise* and assumes that the class labels may be pervaded by random errors. In this case, class labels are thus *precise*, but *uncertain*. Recent contributions along these lines can by found in Refs. [15–17]. In the first two papers, a generative model of label noise is assumed. It is then proposed to model the label noise process by conditional distributions specifying the probabilities that samples labelled as belonging to one class, were in fact drawn from another class. The parameters of such model are then learnt by maximizing the likelihood of the observations knowing the labels [15] or are optimized using a classification maximum likelihood approach [16]. A kernelized version of this kind of approach has been proposed in Refs. [15,18].

The investigations reported in this paper provide a solution to deal with imprecise and/or uncertain class labels, and can therefore be seen as addressing a more general issue than in the above paradigms. Our approach is based on the theory of belief functions [19,20], a framework known to be well suited to represent imprecise and uncertain information. In this paper, we explore its use to represent knowledge on class membership of learning examples, in order to extend the partially supervised framework. In this way, both the uncertainty and the imprecision of class labels may be handled. The considered training sets are of the form $\mathbf{X}^{iu}=\{(\mathbf{x}_1,m_1),\ldots,(\mathbf{x}_N,m_N)\}$, where m_i is a basic belief assignment (bba), or Dempster–Shafer mass function [19] encoding our knowledge about the class of example i. The m_i's (hereafter referred to as "soft labels") may represent different kinds of knowledge, from precise to imprecise and from certain to uncertain. Thus, previous problems are special cases of this general formulation. Other studies have already proposed solutions in which class labels are expressed by possibility distributions or belief functions [21–24]. These labels are interesting when they are supplied by one or several experts and when crisp assignments are hard to obtain. In such cases, the elicitation of experts' opinions regarding the class membership of objects under consideration, in term of possibility or belief functions, can be of interest [21,25].

In this article, we present a new approach to solve learning problems of this type, based on a preliminary study by Vannoorenberghe and Smets [26,27]. This solution is based on mixture models, and therefore assumes a generative model for the data. Generative models have already proved their efficiency in a lot of applications [28]. Their flexibility offers also a good way to benefit from domain specific knowledge, as shown, for example, in text classification [29]. Finally, the adaptability of the expectation maximization (EM) algorithm, which may easily handle specific constraints, is an advantage of generative models. Note that the approach introduced in Refs. [26,30] to apply the EM algorithm to data with soft labels, although based on strong intuitions, was only imperfectly formalized. It was not clear, in particular, what was the equivalent of the log-likelihood function in this case, and if the proposed extension of the EM algorithm converged at all. Precise answers to these questions are provided here.

This article is organized as follows. Background material on belief functions and on the estimation of parameters in mixture models using the EM algorithm will first be recalled in Sections 2 and 3, respectively. The problem of learning from data with soft labels will then be addressed in Section 4, which constitutes the core of the paper. A criterion extending the usual likelihood criterion will first be derived in Section 4.1, and a version of the EM algorithm that optimizes this criterion will be introduced in Section 4.2. Practical considerations and a general discussion will be presented in Sections 4.3 and 4.4, respectively. Finally, simulation results illustrating the advantages of this approach will be reported in Section 5, and Section 6 will conclude the paper.

2. Background on belief functions

2.1. Belief functions on a finite frame

The theory of belief functions was introduced by Dempster [31] and Shafer [19]. The interpretation adopted throughout this paper will be that of the transferable belief model (TBM) introduced by Smets [20]. The first building block of belief function theory is the *bba*, which models the beliefs held by an agent regarding the actual value of a given variable taking values in a finite domain (or *frame of discernment*) Ω, based on some body of evidence. A bba m^{Ω} is a mapping from 2^{Ω} to $[0,1]$ verifying

$$\sum_{\omega\subseteq\Omega} m^{\Omega}(\omega)=1. \tag{1}$$

Each mass $m^{\Omega}(\omega)$ is interpreted as the part of the agent's belief allocated to the hypothesis that the variable takes some value in ω [19,20]. The subsets ω for which $m^{\Omega}(\omega)>0$ are called the *focal sets*. A *categorical* bba has only one focal set. A *simple* bba has at most two focal sets, including Ω. A *Bayesian* bba is a bba whose focal sets are singletons. A bba is said to be *consonant* if its focal sets are nested.

A bba is in one to one correspondence with other representations of the agent's belief, including the plausibility function defined as

$$pl^{\Omega}(\omega)\triangleq\sum_{\alpha\cap\omega\neq\emptyset} m^{\Omega}(\alpha),\quad \forall\omega\subseteq\Omega. \tag{2}$$

The quantity $pl^{\Omega}(\omega)$ is thus equal to the sum of the basic belief masses assigned to propositions that are not in contradiction with ω; it corresponds to the maximum degree of support that could be given to ω, if further evidence become available. The plausibility function associated with a Bayesian bba is a probability measure. If m^{Ω} is consonant, then pl^{Ω} is a possibility measure: it verifies $pl^{\Omega}(\alpha\cup\beta)=\max(pl^{\Omega}(\alpha),pl^{\Omega}(\beta))$, for all $\alpha,\beta\subseteq\Omega$.

2.2. Conditioning and combination

Given two bbas $m_1^{\Omega}, m_2^{\Omega}$ supported by two distinct bodies of evidence, we may build a new bba $m^{\Omega}_{1\textcircled{\cap}2}=m_1^{\Omega}\textcircled{\cap}m_2^{\Omega}$ that corresponds to the conjunction of these two bodies of evidence as

$$m^{\Omega}_{1\textcircled{\cap}2}(\omega)\triangleq\sum_{\alpha_1\cap\alpha_2=\omega} m_1^{\Omega}(\alpha_1)m_2^{\Omega}(\alpha_2),\quad \forall\omega\subseteq\Omega. \tag{3}$$

This operation is usually referred to as the *unnormalized Dempster's rule*, or the *TBM conjunctive rule*. Any positive mass assigned to the empty set during the combination process is interpreted as indicating partial conflict between the two bodies of evidence. If the frame of discernment is supposed to be exhaustive, this mass is usually reallocated to other subsets, leading to the definition of the normalized Demspter's rule $\oplus$ defined as

$$m^{\Omega}_{1\oplus2}(\omega)=\begin{cases}0 & \text{if } \omega=\emptyset,\\ \dfrac{m^{\Omega}_{1\textcircled{\cap}2}(\omega)}{1-m^{\Omega}_{1\textcircled{\cap}2}(\emptyset)} & \text{if } \omega\subseteq\Omega,\omega\neq\emptyset,\end{cases} \tag{4}$$

which is well defined provided $m^{\Omega}_{1\textcircled{\cap}2}(\emptyset)\neq1$. Note that, if m_1^{Ω} (or m_2^{Ω}) is Bayesian, then $m^{\Omega}_{1\oplus2}(\omega)$ is also Bayesian.

The combination of a bba m^{Ω} with a categorical bba focused on $\alpha\subseteq\Omega$ using the TBM conjunctive rule is called (unnormalized) *conditioning*. The resulting bba is denoted $m^{\Omega}(\omega|\alpha)$. Probabilistic conditioning is recovered when m^{Ω} is Bayesian, and normalization is

Download English Version:

https://daneshyari.com/en/article/533684

Download Persian Version:

https://daneshyari.com/article/533684

Daneshyari.com