International Journal of Approximate Reasoning 72 (2016) 43–54

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Active classification using belief functions and information gain maximization

Thomas Reineking

Cognitive Neuroinformatics, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 30 June 2015
Accepted 9 December 2015
Available online 14 December 2015

Keywords:
Belief functions
Classification
Information gain
Object recognition

ABSTRACT

Obtaining reliable estimates of the parameters of a probabilistic classification model is often a challenging problem because the amount of available training data is limited. In this paper, we present a classification approach based on belief functions that makes the uncertainty resulting from limited amounts of training data explicit and thereby improves classification performance. In addition, we model classification as an active information acquisition problem where features are sequentially selected by maximizing the expected information gain with respect to the current belief distribution, thus reducing uncertainty as quickly as possible. For this, we consider different measures of uncertainty for belief functions and provide efficient algorithms for computing them. As a result, only a small subset of features need to be extracted without negatively impacting the recognition rate. We evaluate our approach on an object recognition task where we compare different evidential and Bayesian methods for obtaining likelihoods from training data and we investigate the influence of different uncertainty measures on the feature selection process.

1. Introduction

Probabilistic classification approaches are very popular and one of the most fundamental problems when applying them is to obtain good approximations of the underlying probability distributions. In case there is not enough training data with respect to the number of model parameters, the probability approximations can become quite poor due to overfitting and generalization performance suffers as a consequence. Here belief functions [1,2] provide an interesting alternative to Bayesian methods because belief functions can make the lack of evidence caused by limited amounts of training data explicit, causing the classification to be more robust.

In this paper, we propose a classification approach which combines belief function inference with an active selection of features based on maximizing the expected information gain.[1] Belief functions are used for combining the collected evidence over time while taking into account the amount of available training data for each class. This evidential approach can be viewed as an alternative to Bayesian methods which require specifying a prior on the parameters of the classification model [4]. However, choosing an adequate prior is often quite difficult and the choice strongly affects the resulting classification performance. This requirement is not present for belief functions, but in case a Bayesian prior is available, it can be used, in which case the evidential approach reduces to a Bayesian one.

E-mail address: reineking@uni-bremen.de.

[1] This paper is a revised and extended version of the work presented in [3].

http://dx.doi.org/10.1016/j.ijar.2015.12.005
0888-613X/

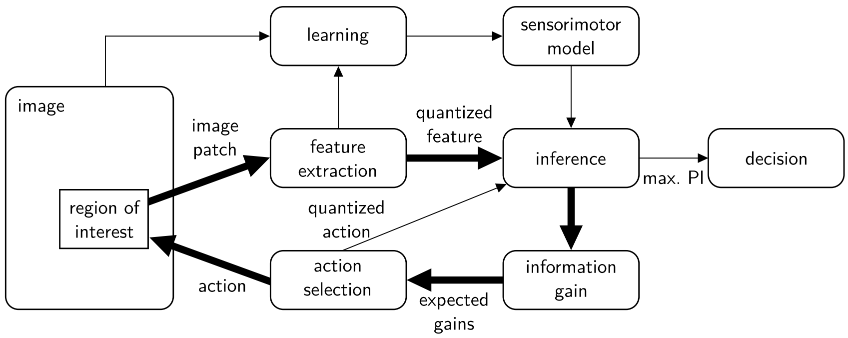

Fig. 1. Schematic overview of the active classification process. The continuous evidence selection and feature extraction cycle is indicated by the thick arrows. An image region is selected based on the current belief (computed in the inference step) and the information gain strategy. The newly acquired evidence is then used to update the current belief which, in turn, leads to another information-gathering action. The procedure stops once the belief is sufficiently unambiguous or all features have been extracted. Figure adapted from [3].

We use an active classification approach where, at each point in time, an information-gathering "action" is selected and executed which leads to the extraction of a new feature. The action selection is based on the principle of maximum expected information gain, which means that the expected uncertainty of the updated belief distribution over possible object classes is reduced as much as possible. This results in a continuous cycle of feature extraction and action selection which is illustrated in Fig. 1.

Reducing the expected uncertainty for active classification was also proposed in [5] in the context of probabilistic classification models, where a set of base classifiers is dynamically selected for each test sample. This is in contrast to other popular ensemble approaches like boosting [6], which use a static combination of classifiers determined during training. Uncertainty reduction is furthermore a common approach in robotics for problems like active localization and exploration [7,8]. Aside from reducing uncertainty, some active classification approaches instead minimize the expected classification error [9,10], which leads to similar results.

In order to compute the expected information gain of an action, the joint distribution of actions and features needs to be modeled. However, the joint distribution is also useful for the classification itself because actions are not just means for acquiring new evidence/features but also provide additional evidence in conjunction with extracted features. In the case of object recognition, this joint distribution contains information about the typical locations of features within an image and utilizing this information generally increases the recognition rate [11].

The next section presents the classification approach and shows how belief functions can be used to model and combine all the evidence. The information gain maximization is discussed in Sect. 3 where two different measures of uncertainty for belief functions are considered. Different methods from the literature for obtaining likelihoods from limited amounts of training data are presented in Sect. 4. In Sect. 5, the classification approach is applied to an object recognition task where the different likelihood computation methods are compared empirically. The paper concludes with a discussion of the presented approach in Sect. 6.

2. Evidential classification

We use a generative evidential classification model [12,13] where the generalized Bayesian theorem [14] is used to compute the posterior class distribution from the likelihoods induced by extracted features. Let $\mathcal{C} = \{c_1, \dots, c_n\}$ denote the frame of discernment for the classification where each c_i represents a class. Furthermore, let $s^{(1:t)} = s^{(1)}, \dots, s^{(t)}$ denote the sequence of features collected up to time t and let $a^{(1:t)}$ denote the sequence of performed actions that resulted in these features.[2] In the case of image classification, an action simply represents an image location but in other settings actions could be more complex (e.g. motor actions of a mobile robot during self localization). We refer to the tuple $(s^{(j)}, a^{(j)})$ as the "sensorimotor feature" at time j. The corresponding frames of discernment $\mathcal{S}$ and $\mathcal{A}$ with $s^{(j)} \in \mathcal{S}$ and $a^{(j)} \in \mathcal{A}$ are finite because we apply vector quantization to both image features and actions.

Because modeling the joint distribution of all sensorimotor features over time without restrictions is infeasible, sensorimotor features are assumed to be conditionally independent given the class [2, Ch. 7], thus resembling a Naive Bayes model [15]. While this is a strong simplification, Naive Bayes models work surprisingly well in practice even if conditional independence is clearly violated [16,17]. The likelihood of class $c_i \in \mathcal{C}$ at time t, which we denote by $L_t(c_i)$, is the plausibility of

[2] We use subscripts to denote elements within a set (e.g. $c_i \in \mathcal{C}$) and superscripts to denote time (e.g. $s^{(t)} \in \mathcal{S}$).