Theoretical Computer Science 589 (2015) 61–75

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

A probabilistic approach to case-based inference

Martin Anthony[a], Joel Ratsaby[b,*]

[a] Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A2AE, UK
[b] Electrical and Electronics Engineering Department, Ariel University, Ariel 40700, Israel

ARTICLE INFO

Article history:
Received 21 February 2013
Received in revised form 11 April 2015
Accepted 14 April 2015
Available online 20 April 2015
Communicated by R. Gavaldà

Keywords:
Case based learning
Multi-category classification
Generalization error
Machine learning
Pattern recognition

ABSTRACT

The central problem in case based reasoning (CBR) is to infer a solution for a new problem-instance by using a collection of existing problem–solution cases. The basic heuristic guiding CBR is the hypothesis that similar problems have similar solutions. Recently, some attempts at formalizing CBR in a theoretical framework have been made, including work by Hüllermeier who established a link between CBR and the probably approximately correct (PAC) theoretical model of learning in his 'case-based inference' (CBI) formulation. In this paper we develop further such probabilistic modelling, framing CBI it as a multi-category classification problem. We use a recently-developed notion of geometric margin of classification to obtain generalization error bounds.

1. Introduction and related work

The basic problem in case based reasoning (CBR) is to infer a solution for a new problem-instance by using a collection of existing problem–solution cases [1]. (We will henceforth use 'problem' for 'problem instance'.) The basic heuristic that guides CBR is the hypothesis that similar problems have similar solutions (see [2], for example). The area of CBR research has had practical success and has been shown to be widely applicable [3]. The well known methodological framework of case-based reasoning divides CBR into four main steps (referred to as the R^4 framework): retrieve, reuse, refine and retain [2].

There have been a number of attempts to develop a sound theoretical basis for CBR. Significant recent work, due to Hüllermeier [4], makes a connection between CBR and the probably approximately correct (PAC) theoretical model of learning [5]. Hüllermeier defines case-based reasoning as a prediction process, which allows him to make the connection between CBR and the learning based on a sample. He calls this framework *case-based inference* (CBI) and it aims to solve the 'retrieve' and 'reuse' steps of the R^4 framework. Given a new problem to be solved, CBI aims just to produce a 'promising' set of solutions for use by the remaining two steps of the R^4 framework. The last two stages of the R^4 framework use not just the set of candidate solutions but also domain-knowledge, user input and further problem-solving strategies [2]. As noted in Section 5.4 of [6], these steps *adapt* the set of promising solutions into a solution that fits the existing problem.

In this paper, we continue work in the direction inspired by [2], probabilistically modelling case-based inference as a multi-category classification problem. We use a recently-developed notion of geometric margin of classification, called width, to obtain generalization error bounds. This notion has recently been used in [7] to exploit regularity in training

[*] Corresponding author.
E-mail addresses: m.anthony@lse.ac.uk (M. Anthony), ratsaby@ariel.ac.il (J. Ratsaby).

http://dx.doi.org/10.1016/j.tcs.2015.04.016
0304-3975/© 2015 Elsevier B.V. All rights reserved.

samples for the problem of classification learning in finite metric spaces. The main results in the current paper are bounds on the error of case-based learning which involve the sample width.

Dubois and Prade [8] and Dubois et al. [9] attempted to provide a formal model of CBR which is based on fuzzy logic. The similarity between two problems (or two solutions) is represented by a fuzzy relations. There is no learning process for determining these relations. Our model differs from theirs in that we learn from examples to produce a set of candidate solutions for input problems; and we do not employ fuzzy logic, but statistical learning under the PAC framework.

Ontañón and Plaza [10] introduce a model of knowledge transfer for case-based inference part of CBR. It produces, from retrieved cases (cases whose problems are similar to the given problem), a set of conjectures (or incomplete solutions) rather than actual solutions. A conjecture may require further adaptation, for instance using some domain specific rules, in order to produce a solution. (Their model can deal with cases where there is no clear distinction between a problem and solution.) Our model differs from theirs in that it produces a set of complete solutions for the input problem (rather than conjectures); and our model expands on the CBI framework of Hüllermeier, and hence we have two separate spaces, one for solutions and one for problems.

In Section 2, we start by describing Hüllermeier's framework of CBI, where the goal is to predict a 'credible' or promising set of solutions for a given input problem instance. We outline and explain our contribution and its connections with this framework. The key idea is that we model CBI as a supervised learning problem. Section 3 describes a probabilistic model that is the basis of our analysis. We redefine what is meant by a credible set in this context and we provide a mathematical formalism for measuring the success of a method for predicting credible sets. Section 4 presents some recent results on the generalization accuracy of learning multi-category classifiers defined on metric spaces, and provides results on which we draw for the conclusions of this paper. Section 5 describes in detail the important transformation of learning CBI to the problem of supervised learning. Section 6 provides bounds on the error of learning CBI. These bounds can serve as a guiding criterion for the design of successful algorithms.

One main contribution is to show how learning CBI over the wide spectrum of complex and unstructured CBR domains can be transformed to standard supervised learning problems. A further contribution is in showing how the large-width advantage (familiar from the branch of learning theory known as large-margin learning) can also be realised for learning CBI.

2. Case-based inference (CBI)

In the Introduction, we mentioned that CBI infers as an output a set of candidate, or 'promising', solutions rather than solving the full CBR problem by predicting a single specific solution. This is at the basis of what Hüllermeier [4] calls *approximate reasoning*. We now describe his framework (using slightly different notation).

2.1. Hüllermeier's CBI framework

In the general set-up of Hüllermeier's case-based inference, there is a problem space, denoted by $\mathcal{X}$, and a solution space, denoted by $\mathcal{Y}$. We define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The problem space and solution space may be very general; in particular, not only finite-dimensional vector spaces (as those that are common in supervised learning) but also problems described by more complex structures like trees, graphs, or plans. Each of the spaces, $\mathcal{X}$, $\mathcal{Y}$, has a similarity function, $\text{sim}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to [0,1]$ and $\text{sim}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, respectively. These are reflexive and symmetric; that is, $\text{sim}_{\mathcal{X}}(x,x) = 1$, and $\text{sim}_{\mathcal{X}}(x,x') = \text{sim}_{\mathcal{X}}(x',x)$, and similarly for $\text{sim}_{\mathcal{Y}}$. The goal of case-based inference in [4] can be described as follows.

Goal of CBI in [4]: Given a sample $\{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ (also referred to as a case-base), consisting of problem–solution pairs, and given a new problem instance x, produce for it a subset of solutions (subset of $\mathcal{Y}$) called a *credible set*, that contains some (possibly all) solutions for the problem x.

An underlying assumption is that there exists some unknown relationship between the level of similarity of pairs of problems and the similarity of their solutions. Hüllermeier [4] represents this by a *similarity profile* σ, mapping from $[0,1]$ to $[0,1]$ and defined by

$$\sigma(\alpha) := \inf_{x,x' \in \mathcal{X} : \text{sim}_{\mathcal{X}}(x,x') = \alpha} \text{sim}_{\mathcal{Y}}(y, y')$$

where (x,y), $(x',y') \in \mathcal{Z}$ are two problem–solution pairs. This function σ represents in a formal way the CBR assumption that similar problems have similar solutions, since given any pair of problems that are similar by a value of α, their solutions must be at least similar by a level of $\sigma(\alpha)$.

Theoretically speaking, if one knows σ then, for a given problem x, one can produce a 'credible' set of solutions, which is defined as

$$C(x) := \bigcap_{i=1}^{m} \Gamma_\sigma(z_i, x) \tag{1}$$