

Complexity-Guided Case Discovery for Case Based Reasoning

Stewart Massie and Susan Crow and Nirmalie Wiratunga

The Robert Gordon University
Aberdeen, AB25 1HG, Scotland, UK

sm@comp.rgu.ac.uk smc@comp.rgu.ac.uk nw@comp.rgu.ac.uk

Abstract

The distribution of cases in the case base is critical to the performance of a Case Based Reasoning system. The case author is given little support in the positioning of new cases during the development stage of a case base. In this paper we argue that classification boundaries represent important regions of the problem space. They are used to identify locations where new cases should be acquired. We introduce two complexity-guided algorithms which use a local complexity measure and boundary identification techniques to actively discover cases close to boundaries. The ability of these algorithms to discover new cases that significantly improve the accuracy of case bases is demonstrated on five public domain classification datasets.

Introduction

Case Based Reasoning (CBR) solves new problems by re-using the solution of previously solved problems. The case base is the main source of knowledge in a CBR system and, hence, the availability of cases is crucial to a system's performance. It is the availability of cases that often supports the choice of CBR for problem-solving tasks, however, in real environments there are often gaps in the coverage of the case base because it is difficult to obtain a collection of cases to cover all problem-solving situations.

Adaptation knowledge can be used to provide solutions to new problems that occur in the gaps that result from a lack of case coverage. However, gaining effective adaptation knowledge may be impossible or require considerable knowledge acquisition effort. The inclusion of additional, strategically placed cases can provide a more cost-effective solution.

Case discovery is the process of identifying *useful* new cases to fill gaps that exist in the coverage of the case base. This is different from traditional case learning, through the retain stage of the CBR cycle, in which newly solved problems are routinely added to the case base to assist future problem-solving. Rather case discovery is an active learning problem in which the aim is to identify areas of the problem space in which new cases would help to improve the system's performance and to create cases to fill these gaps. Commercial systems generally assume that a suitable case

base already exists and give little help to the case author in the case discovery stage of building the case base. There is a need for techniques to assist the case author during this crucial case base development stage.

We argue that new cases should be placed in regions of the problem-space in which the system is uncertain of the solution, and that these regions are generally close to boundaries between classifications. In this paper we present a new technique to identify and rank these areas of uncertainty and create candidate cases to assist the case author place new cases in these regions.

The remainder of this paper describes our approach and evaluates it on several public domain case bases. The next section discusses existing work on case discovery. The following sections outline how we use a complexity metric, boundary detection and clustering to identify areas of the problem-space that need the support of new cases and how these cases are created. The approach is then evaluated against two benchmark algorithms before we draw some final conclusions.

Related Work in Case Discovery

The case discovery problem can be considered in two stages. First *interesting* areas or gaps within the coverage of the case base must be identified and secondly cases must be created to fill these gaps. This presents a more complex challenge when compared to the more commonly researched case base editing or selective sampling problems that have a pool of existing cases from which to select cases. In contrast, the task of case discovery is to add to the case knowledge using implicit information held within the case base.

Some research has focused on the first stage of the discovery process. One approach to identifying gaps has been to focus on locating maximal empty hyper-rectangles within k -dimensional space (Liu, Ku, & Hsu 1997). In their later research the algorithm is capable of locating hyper-rectangles within data containing both continuous and discrete valued attributes (Liu *et al.* 1998). The main problem with this approach is that there is no way to identify if the gap found in the problem space is *interesting* from a problem-solving view-point, or even represents a possible combination of attribute values. An alternative approach to identifying interesting areas for new cases is proposed in (Wiratunga, Crow, & Massie 2003) as part of a selective sampling technique.

