Abstract. Inductive logic programming (ILP) is built on a foundation laid by research in other areas of computational logic. But in spite of this strong foundation, at 10 years of age ILP now faces a number of new challenges brought on by exciting application opportunities. The purpose of this paper is to interest researchers from other areas of computational logic in contributing their special skill sets to help ILP meet these challenges. The paper presents five future research directions for ILP and points to initial approaches or results where they exist. It is hoped that the paper will motivate researchers from throughout computational logic to invest some time into “doing” ILP.

1 Introduction

Inductive Logic Programming has its foundations in computational logic, including logic programming, knowledge representation and reasoning, and automated theorem proving. These foundations go well beyond the obvious basis in definite clause logic and SLD-resolution. In addition ILP has heavily utilized such theoretical results from computational logic as Lee’s Subsumption Theorem [18], Gottlob’s Lemma linking implication and subsumption [12], Marcinkowski and Pacholski’s result on the undecidability of implication between definite clauses [22], and many others. In addition to utilizing such theoretical results, ILP depends crucially on important advances in logic programming implementations. For example, many of the applications summarized in the next brief section were possible only because of fast deductive inference based on indexing, partial compilation, etc. as embodied in the best current Prolog implementations. Furthermore, research in computational logic has yielded numerous important lessons about the art of knowledge representation in logic that have formed the basis for applications. Just as one example, definite clause grammars are central to several ILP applications within both natural language processing and bioinformatics.

ILP researchers fully appreciate the debt we owe to the rest of computational logic, and we are grateful for the foundation that computational logic has provided. Nevertheless, the goal of this paper is not merely to express gratitude, but
also to point to the present and future needs of ILP research. More specifically, the goal is to lay out future directions for ILP research and to attract researchers from the various other areas of computational logic to contribute their unique skill sets to some of the challenges that ILP now faces. In order to discuss these new challenges, it is necessary to first briefly survey some of the most challenging application domains of the future. Section 2 provides such a review. Based on this review, Section 3 details five important research directions and concomitant challenges for ILP, and Section 4 tries to “close the sale” in terms of attracting new researchers.

2 A Brief Review of Some Application Areas

One of the most important application domains for machine learning in general is bioinformatics, broadly interpreted. This domain is particularly attractive for (1) its obvious importance to society, and (2) the plethora of large and growing data sets. Data sets obviously include the newly completed and available DNA sequences for *C. elegans* (nematode), *Drosophila* (fruitfly), and (depending on one’s definitions of “completed” and “available”) man. But other data sets include gene expression data (recording the degree to which various genes are expressed as protein in a tissue sample), bio-activity data on potential drug molecules, x-ray crystallography and NMR data on protein structure, and many others. Bioinformatics has been a particularly strong application area for ILP, dating back to the start of Stephen Muggleton’s collaborations with Mike Sternberg and Ross King [29, 16]. Application areas include protein structure prediction [29, 37], mutagenicity prediction [17], and pharmacophore discovery [7] (discovery of a 3D substructure responsible for drug activity that can be used to guide the search for new drugs with similar activity). ILP is particularly well-suited for bioinformatics tasks because of its abilities to take into account background knowledge and structured data and to produce human-comprehensible results. For example, the following is a potential pharmacophore for ACE inhibition (a form of hypertension medication), where the spacial relationships are described through pairwise distances.

Molecule A is an ACE inhibitor if:

molecule A contains a zinc binding site B, and
molecule A contains a hydrogen acceptor C, and
the distance between B and C is 7.9 +/- 0.75 Angstroms, and
molecule A contains a hydrogen acceptor D, and
the distance between B and D is 8.5 +/- 0.75 Angstroms, and
the distance between C and D is 2.1 +/- 0.75 Angstroms, and
molecule A contains a hydrogen acceptor E, and
the distance between B and E is 4.9 +/- 0.75 Angstroms, and

1 Not to put too fine a point on the matter, this paper contains unapologetic proselytizing.

2 Hydrogen acceptors are atoms with a weak negative charge. Ordinarily, zinc-binding would be irrelevant; it is relevant here because ACE is one of several proteins in the body that typically contains an associated zinc ion. This is an automatically generated translation of an ILP-generated clause.