Learning under uniform distribution

A.Marchetti-Spaccamela
Dept. of Mathematics
University of L'Aquila
L'Aquila, Italy

M.Protasi
Dept. of Mathematics
University of Roma "Tor Vergata"
Roma, Italy

Abstract We study the learnability from examples of boolean formulae assuming that the examples satisfy a uniform distribution assumption. We analyze the requirements of known algorithms (upper and lower bounds) under uniform distribution and we propose a new combinatorial measure in order to characterize the complexity of boolean formulae.

1 Introduction

Algorithms for learning boolean formulae have been widely studied in the last years since Valiant's paper [V84] that showed how it is possible to study learning algorithms in the framework of computational complexity theory. In this approach we assume that the learner has access to data that exemplify the formula to be learned in a positive or in a negative way and that he has performed its task when he can find a good approximation of the formula with sufficiently large probability.

One of the most important classes from the learning point of view is the class DNF , that is the class of formulae in disjunctive normal form; however, since the class DNF is not efficiently learnable, several subclasses have been investigated. Let k-DNF (k-CNF) be the class of boolean formulae in disjuntive (conjunctive) normal form where each term (clause) is the

Work partially supported by Italian project MPI 'Algoritmi e Strutture di Calcolo'.

product (sum) of at most k literals. Let d-term-DNF (d-clause-CNF) be the class of boolean formulae in disjuntive (conjunctive) normal form with at most d terms. Furthermore, let k-d-term-DNF (k-d-clause-CNF) be the subclass of k-DNF (k-CNF) formulae with at most d terms (clause). A formula f is monotone if it has not negated variables.

The majority of the results have been given assuming that the examples are drawn according to a probability distribution which is fixed but unknown (distribution free model). However, because of the generality of the approach, few classes can be learned and many negative results occur. For example in [KLPV87] it has been shown that the problem of learning monotone d-term DNF by d-term DNF formulae ($d \geq 2$) is NP-hard. There is also a stronger result; namely it is hard to learn a formula even if we allow the formula to have about twice as many disjuncts as the formula to be learned. Formally, if $d > 5$ then monotone d-term DNF are not learnable by (2 d - 5)-term-DNF formulae.

A further limitation of the distribution free model is that even if the number of examples is polynomially bounded it can be too large from a practical point of view. For these reasons it is useful to restrict the class of probability distribution used for generating the examples. The case in which the uniform distribution is considered has been previously considered by several authors [KLPV87], [BI88], [KMP88]; in this paper we continue this approach. The aim is to give stronger and more efficient results for some classes of formulae. More precisely we are mainly interested in studying the sample complexity, that is the number of examples required.

In Section 3 we analyse Valiant's algorithm for learning k-DNF formulae in the case of k-d-term-DNF formulae and we show that, for formulae with small k and d, the use of uniform distribution allows to increase the efficiency of the algorithm. On the other side we show that for a whole class of algorithms the performance obtained in the distribution free model cannot be improved significantly. In Section 4 we introduce the notion of term similarity of a formula. This notion is an attempt towards the aim of characterizing the performance of a learning algorithm from a combinatorial point of view under uniform distribution and we show how term similarity affects the complexity of learning process in the case of the k-d-term-DNF formulae. In Section 5, we concentrate on d-term-DNF formulae; we prove a lower bound on the number of examples for the class of algorithms studied in Section 3. Furthermore we study how the uniform distribution affects the sample complexity of an algorithm introduced in [KMP88] for DNF formulae.

2 Basic definitions

First of all we give some standard definitions. Let n be a natural number. A concept F is a boolean function with domain $\{0,1\}^n$. Given a vector X, X is a positive example of the concept F, if $F(X) = 1$, otherwise X is a negative example.