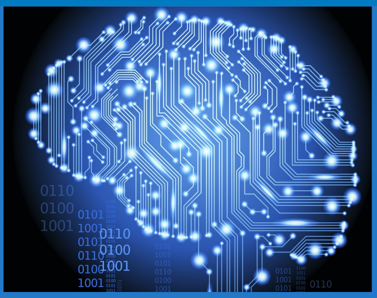
Department of Machine Learning



Institute of Computer Science, CAS Prague

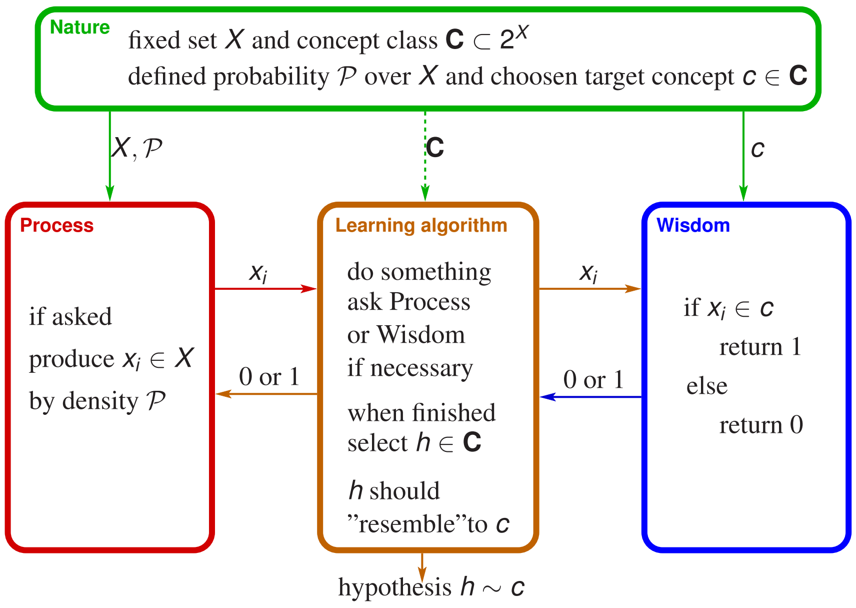
Estimates of the number of patterns in the PAC model for nonconsistent separation.

František Hakl

Topic Description

Separation methods have been used in process management and decision-making very often. A relatively frequent case, posed by different practical situations, is the necessity of deciding whether a vector of state variables describing a real natural or technical system does or does not belong to a set that may be classified as critical, i.e. corresponding to a state in which the system under observation should not appear. Separation algorithms are very often based on the knowledge of the previous history of the given system, while the describing vectors for both states from the critical set as well as those from the complementary set are known. Seen in this light, separation algorithm has at its disposal a set of states and, furthermore, information saying to which particular set these states belong. The purpose of separation algorithm is to attempt, on the basis of this knowledge, to set its inner parameters in order to perform required separation in the requested quality. This particular mode of assigning a separation task is called supervised learning with membership queries. The question is whether we can, in any way, evaluate the credibility of separation, i.e. anticipation that the new, hitherto unknown vector of state variables will be assigned to the correct subset of states of the given system. We would like to perform this evaluation of credibility prior to using separation algorithm in an effort to refrain from employing other verification monitoring of system, which – in case of critical states – understandably bring undesirable or unacceptable phenomena, eventually high costs. For this purpose, the entire separation process has been formalized in the PAC (Probably Approximately Correct) learning model. This – assisted by the term of the Vapnik-Chervonenkis dimension of the system of subsets – derives the necessary size of the set of known states, used during the learning process that guarantees a small degree of probability ($< \epsilon$) of committing serious mistakes ($> \delta$) in assigning state vectors to the correct subset. Such a separation algorithm is called (ϵ, δ)-learning algorithm.

Basic Diagram of the PAC Model



Nature

fixed set X and concept class $\mathbf{C} \subset 2^X$

defined probability $\mathcal{P}$ over X and choosen target concept $c \in \mathbf{C}$

$X, \mathcal{P}$

$\mathbf{C}$

c

Process

if asked

produce $x_i \in X$

by density $\mathcal{P}$

x_i

0 or 1

Learning algorithm

do something
ask Process
or Wisdom
if necessary

when finished
select $h \in \mathbf{C}$

h should
"resemble"to c

x_i

0 or 1

Wisdom

if $x_i \in c$
return 1
else
return 0

hypothesis $h \sim c$

Estimates of the Number of Patterns in the Standard PAC Model

Let $\mathbf{C}$ be a nontrivial concept class. Then:

1. If $\mathtt{VCdim}(\mathbf{C}) = d < +\infty$, then:
   - for arbitrary $0 < \epsilon < \frac{1}{2}$ does not exist (ϵ, δ)-learning algorithm exploiting less than

     $$\max\left(\frac{1-\epsilon}{\epsilon}\log_e\left(\frac{1}{\delta}\right), d\left(1 - 2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)\right) \tag{1}$$

     queries.
   - for arbitrary $0 < \epsilon < 1$, any learning algorithm producing always consistent hypothesis and exploiting at least

     $$\max\left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right) \tag{2}$$

     queries is (ϵ, δ)-learning algorithm.
2. For $\mathbf{C}$ there exists (ϵ, δ)-learning algorithm if and only if $\mathtt{VCdim}(\mathbf{C}) < +\infty$.

Practical Applicability of the PAC Model

In the standard PAC model, estimates of the number of queries of the learning algorithm can be performed if the two assumptions given below are valid:

- the so-called consistency, which assumes that the learning algorithm does not make mistakes in the hitherto known reality, i.e. that it accurately classifies state vectors used during learning. Unfortunately, in the separation methods used in real situations, this fact is guaranteed only very rarely indeed.
- disjunction of the critical set and the state vectors describing non-critical states. Neither this fact is complied with in real situations; de facto, the incidence of state vectors in both separated sets is given by probability densities whose carriers (spheres of non-zero probability) are not disjunctive.

These two facts mean that in a majority of practical cases, given the previously required accuracy of separation, estimates of the adequate number of queries cannot be applied. But the substance of the issue implies that a small deviation from the above-mentioned assumptions tends to lead to a minor disruption of basic estimates of the number of queries. The subject of this work is to study this particular dependence with the aim of modifying the theory of the PAC model in a way to incorporate the rate of non-compliance with the above assumptions for consistency and disjunction.

Topics You Could Address

- Elaboration of the modes of quantifying non-compliance with the conditions of consistency and disjunction.
- Derivation of theoretical formulas of the PAC model for estimation of the number of queries reflecting the rate of disruption of the assumptions concerning the standard PAC model.
- Comparison whether the derived estimates match the results of the separation of selected real-life data.

http://ml.cs.cas.cz

hakl@cs.cas.cz