Electronic Colloquium on Computational Complexity, Report No. 66 (2000)

On Learning from Ambiguous Information

Peter Auer*

Institute for Theoretical Computer Science, Graz University of Technology

Klosterwiesgasse 32/2, A-8010 Graz, Austria

Email: pauer@igi.tu-graz.ac.at

Fax: +43 316 873-5805. Phone: +43 316 873-5821

June 30, 1997

Abstract

We investigate a variant of the Probably Almost Correct learning model where the learner has to learn from ambiguous information. The ambiguity is introduced by assuming that the learner does not receive single instances with their correct labels as training data, but that the learner receives tuples of instances where a tuple has a negative label if *all* instances of the tuple should be labeled as negative and a tuple has a positive label if *at least one* instance of the tuple should be labeled as positive. Thus a positive tuple is ambiguous since it is not known which of its instances is a positive instance.

*Supported by the ESPRIT Project NeuroCOLT.

ISSN 1433-8092

Such ambiguous information is for example relevant in learning problems for drug design. We present an improved algorithm for learning axis-parallel rectangles in this model of ambiguous information. In the drug design domain such rectangles represent the shapes of molecules with certain properties.

Keywords: computational learning theory, classification, multiple instance problem, axis-parallel rectangles.

1 Introduction and statement of results

1.1 The PAC learning model

The PAC learning model was first introduced by [Valiant, 1984]. It gives a formalization of *concept learning* in respect to an underlying distribution D over some domain X. Concepts are modeled as subsets $C \subseteq X$ of the domain X and the class of all relevant concepts is called the *concept class* $\mathcal{C} \subseteq 2^X$. For convenience we sometimes refer to a concept as a function $C : X \to \{+, -\}$ with $C(x) = +$ if $x \in C$ and $C(x) = -$ if $x \notin C$.

The goal of the learner is to calculate a hypothesis $\hat{C} \subseteq X$ which approximates the unknown target concept C to be learned. The quality of the approximation is measured by the underlying distribution D such that $D\{x : C(x) \neq \hat{C}(x)\}$ is the error of hypothesis $\hat{C}$, i.e. the probability that a random instance drawn accordingly to D is incorrectly (in respect to C) classified by $\hat{C}$. If $D\{x : C(x) \neq \hat{C}(x)\} < \epsilon$ we say that $\hat{C}$ as an ϵ-accurate hypothesis.

To obtain a good hypothesis the learner is given a random training sample of labeled instances $(x_1, C(x_1)), \ldots, (x_m, C(x_m))$ drawn independently from D. Since the sample is drawn at random the learner might be unlucky and receive a sample from which one cannot learn much. Thus the learner is required to calculate a good hypothesis only for most of the possible sample draws. Formally, we have the following definition.

Definition 1.1 *An algorithm A PAC-learns the concept class $\mathcal{C} \subseteq 2^X$ with accuracy $\epsilon > 0$ and confidence $\delta > 0$ from m examples if for for all distributions D on X and all $C \in \mathcal{C}$, with probability $1 - \delta$ a random sample of size m is drawn from which algorithm A calculates a hypothesis $\hat{C}$ with $D\{x : C(x) \neq \hat{C}(x)\} < \epsilon$. The input to the algorithm are the parameters ϵ and δ and the random sample.*

A general technique to calculate a good hypothesis is to pick an arbitrary concept $\hat{C} \in \mathcal{C}$ which classifies all examples in the training sample correctly. Then, under some mild conditions and if the training sample is big enough, it can be shown that $\hat{C}$ is a sufficiently accurate hypothesis.

1.2 Ambiguous information — multiple instances

In a variant of the PAC learning model, the multiple instance model, the learner receives a training sample of label r-tuples

$$((x_{1,1}, \ldots, x_{1,r}), \ell_1), \ldots, ((x_{m,1}, \ldots, x_{m,r}), \ell_m)$$

with $\ell_k = +$ if there is a $x_{k,j} \in C$ and $\ell_k = -$ if all $x_{k,j} \notin C$, where C is the target concept. Of course this model could be equivalently embedded into the usual PAC model

by using as domain X^r and as concept class $\mathcal{C}^{(r)} := \{C^{(r)} | C \in \mathcal{C}$ and $(x_1, \ldots, x_r) \in C^{(r)} \Leftrightarrow \exists j : x_j \in C\}$. Unfortunately, it is very likely that in this model even simple concept classes cannot be learned within a reasonable amount of time. Consider for example the class of d-dimensional axis-parallel rectangles[1]

$$\mathcal{C} = \left\{ \prod_{i=1}^{d} [0, b_i] : b_i \geq 0 \right\}.$$

Then there is the following theorem.

Theorem 1.2 ([Auer et al., 1997]) *If the class $\mathcal{C}^{(r)}$ can be learned with arbitrary small accuracy ϵ and confidence δ in time polynomial in d, r, $\frac{1}{\epsilon}$, and $\frac{1}{\delta}$, then $\mathcal{RP} = \mathcal{NP}$.*

The above problem arises since when learning $\mathcal{C}^{(r)}$ the distribution on $\mathbf{R}^{d \times r}$ is arbitrary. To circumvent this problem we assume that all instances $\mathbf{x}_{k,j}$ of a tuple $(\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,r})$ are drawn independently from some distribution D on $\mathbf{R}^d$ ([2]). Thus the distribution on $\mathbf{R}^{d \times r}$ from which the r-tuples are drawn is D^r, i.e. the underlying distributions from which $\mathcal{C}^{(r)}$ has to be learned are restricted to this type. Formally, we have the following model.

Definition 1.3 *An algorithm A learns the concept class of d-dimensional rectangles $\mathcal{C}$ with accuracy $\epsilon > 0$ and confidence $\delta > 0$ from m* independent *r-instance examples if for for all distributions D on $\mathbf{R}^d$ and all $C \in \mathcal{C}$ the following holds: with probability*

[1] We consider only rectangles with their "lower left" corner fixed at $\mathbf{0}$. The generalization to arbitrary axis-parallel rectangles is straight forward.

[2] In the following we denote instances by $\mathbf{x}$ to indicate that they are elements of $\mathbf{R}^d$.

$1-\delta$ *a random sample of* r*-instance examples of size* m *is drawn from* D^r *and labeled by* $C^{(r)}$ *such that algorithm* A *calculates a hypothesis* $\hat{C} \subseteq \mathbf{R}^d$ *with* $D^r\{(\mathbf{x}_1, \ldots, \mathbf{x}_r) : C^{(r)}(\mathbf{x}_1, \ldots, \mathbf{x}_r) \neq \hat{C}^{(r)}(\mathbf{x}_1, \ldots, \mathbf{x}_r)\} < \epsilon$.

1.3 Previous and new results

The first investigation of learning axis-parallel rectangles from multiple instances was undertaken by [Dietterich et al., 1997] in an empirical work for drug design. The first positive theoretical result for learning rectangles from independent multiple instances was obtained by [Long and Tan, 1996], but they had to assume that the underlying distribution is a product distribution on $\mathbf{R}^d$. In [Auer et al., 1997] the restriction to product distributions was removed and the performance bounds were considerably improved.

Theorem 1.4 ([Auer et al., 1997]) *The class of* d*-dimensional axis-parallel rectangles can be learned from*

$$m = O\left(\frac{d^2r^2}{\epsilon^2}\log\frac{d}{\delta}\right)$$

independent r*-instance examples with accuracy* ϵ *and confidence* δ.

A general results for learning from multiple instances was obtained in [Blum and Kalai, 1997]. For the class of rectangles they have the following bound.

Theorem 1.5 ([Blum and Kalai, 1997]) *The class of* d*-dimensional axis-parallel rectangles can be learned from*

$$m = O\left(\frac{d^2r}{\epsilon^2}\left(\log\frac{d}{\delta} + \log\log\frac{r}{\epsilon}\right)\right)$$

independent r-instance examples with accuracy ϵ and confidence δ.

Combining the techniques of [Auer et al., 1997] and [Blum and Kalai, 1997] we are able to obtain an algorithm for learning rectangles from independent multiple instances whose analysis gives performance bounds which improve on both bounds given in Theorems 1.4 and 1.5.

Theorem 1.6 *The class of d-dimensional axis-parallel rectangles can be learned from*

$$m = 38^2 \cdot 32 \cdot \frac{d^2 r}{\epsilon^2} \log \frac{2d+2}{\delta}$$

independent r-instance examples with accuracy ϵ and confidence δ. The run time of the learning algorithm is $O(drm \log m)$.

Remark 1.7 *We did not attempt to optimize the constants.*

2 An improved algorithm for learning rectangles from multiple instances

2.1 Basic idea

The main idea to calculate a good approximation $\hat{C}$ of a target rectangle C is to calculate estimates for $\beta_i(t) = D\{\mathbf{x} : \mathbf{x} \in C \text{ and } x_i > t\}$ which are the probabilities that a random instance is inside the target rectangle and its i-th coordinate is greater than t. Another important quantity is $p = D\{\mathbf{x} : \mathbf{x} \notin C\}$, the probability to draw a random instance outside of the target rectangle. Note that $q = p^r$ is the probability

to draw a negative r-instance example. The following lemmas show that a good approximation $\hat{C}$ of C can be calculated from sufficiently accurate estimates $\hat{\beta}_i(t)$ and $\hat{q}$ of $\beta_i(t)$ and q, respectively.

Lemma 2.1 *If* $|q - \hat{q}| \leq \frac{\epsilon}{4}$ *and* $\hat{q} < \frac{3}{4\epsilon}$ *then* $\hat{C} = \mathbf{R}^d$ *is an* ϵ*-accurate approximation of* C*. Otherwise, if* $|q - \hat{q}| \leq \frac{\epsilon}{4}$ *and* $\hat{q} \geq \frac{3}{4\epsilon}$ *then* $\frac{1}{2} \leq \frac{\hat{q}}{q} \leq \frac{3}{2}$.

Proof. If $\hat{C} = \mathbf{R}^d$ then the hypothesis classifies all examples as positive. Thus the probability of error is $q \leq \hat{q} + |q - \hat{q}| < \epsilon$ if $|q - \hat{q}| \leq \frac{\epsilon}{4}$ and $\hat{q} < \frac{3}{4}\epsilon$. If $\hat{q} \geq \frac{3}{4}\epsilon$ then $q \geq \frac{\epsilon}{2}$ and $|\frac{\hat{q}}{q} - 1| = \left|\frac{\hat{q}-q}{q}\right| \leq \frac{1}{2}$. □

Lemma 2.2 *If* $q \geq \frac{\epsilon}{2}$, $\hat{C} \subseteq C$, *and* $D\{\mathbf{x} : C(\mathbf{x}) \neq \hat{C}(\mathbf{x})\} < \frac{\epsilon p}{3rq}$ *then* $D^r\{(\mathbf{x}_1, \ldots, \mathbf{x}_r) : C^{(r)}(\mathbf{x}_1, \ldots, \mathbf{x}_r) \neq \hat{C}^{(r)}(\mathbf{x}_1, \ldots, \mathbf{x}_r)\} < \epsilon$.

Proof. Since $\hat{C} \subseteq C$

$$
\begin{aligned}
& D^r\{(\mathbf{x}_1, \ldots, \mathbf{x}_r) : C(\mathbf{x}_1, \ldots, \mathbf{x}_r) \neq \hat{C}(\mathbf{x}_1, \ldots, \mathbf{x}_r)\} \\
&= (1 - D\{\hat{C}\})^r - (1 - D\{C\})^r \\
&= (D\{C\} - D\{\hat{C}\}) \sum_{k=1}^{r-1} (1 - D\{C\})^k (1 - D\{\hat{C}\})^{r-k-1} \\
&\leq \frac{\epsilon p}{3q} \left(p + \frac{\epsilon p}{3rq}\right)^{r-1} \leq \frac{\epsilon}{3} \left(1 + \frac{1}{r}\right)^{r-1} \leq \epsilon.
\end{aligned}
$$

□

Lemma 2.3 *If* $|q - \hat{q}| \leq \frac{\epsilon}{4}$, $\hat{q} \geq \frac{3\epsilon}{4}$, *and for all* $i = 1, \ldots, d$ *and all* $t \in \mathbf{R}$, $|\beta_i(t) - \hat{\beta}_i(t)| \leq \frac{\epsilon p}{12drq}$, *then* $\hat{C} = \prod_{i=1}^{d} [0, \hat{b}_i]$,

$$
\hat{b}_i = \inf \left\{ t : \hat{\beta}_i(t) \leq \frac{\epsilon \hat{q}^{1/r}}{8dr\hat{q}} \right\},
$$

is an ϵ*-accurate approximation of* C.

Proof. Let $C = (b_1, \ldots, b_d)$. Then $\beta_i(b_i) = 0$ and $\hat{\beta}_i(b_i) \leq \frac{\epsilon p}{12drq} \leq \frac{\epsilon \hat{q}^{1/r}}{8dr\hat{q}}$ by Lemma 2.1. Thus $\hat{b}_i \leq b_i$. Furthermore $\beta_i(\hat{b}_i) \leq \frac{\epsilon \hat{q}^{1/r}}{8dr\hat{q}} + \frac{\epsilon p}{12drq} \leq \frac{\epsilon p}{3drq}$ since $\beta_i(\cdot)$ is continuous from the right. Hence $\hat{C} \subseteq C$ and $D\{C \setminus \hat{C}\} \leq \sum_{i=1}^{d} \beta_i(\hat{b}_i) \leq \frac{\epsilon p}{3rq}$, and Lemma 2.2 gives the claim. □

2.2 Calculating an estimate for $\beta_i(t)$

To calculate an accurate estimate of $\beta_i(t)$ we introduce the quantities $\alpha_i(t) = D\{\mathbf{x} : x_i > t\}$, the probability of drawing a random instance whose i-th coordinate is greater than t, and $\gamma_i(t) = D\{\mathbf{x} : x_i > t | \mathbf{x} \notin C\}$, the conditional probability of drawing a random instance whose i-th coordinate is greater than t given that the random instance is not inside of the target rectangle. These quantities can be easily estimated from a random sample S of multiple instance examples. Furthermore $\beta_i(t) = \alpha_i(t) - p \cdot \gamma_i(t)$ which yields the estimate

$$\hat{\beta}_i(t) = \hat{\alpha}_i(t) - \hat{p} \cdot \hat{\gamma}_i(t) \tag{1}$$

with

$$\hat{\alpha}_i(t) := \frac{\text{number of instances } \mathbf{x} \text{ in } S \text{ with } x_i > t}{\text{total number of instances in } S}, \tag{2}$$

$$\hat{q} := \frac{\text{number of negative examples in } S}{\text{total number of examples in } S}, \tag{3}$$

$$\hat{p} := \hat{q}^{1/r}, \tag{4}$$

$$\hat{\gamma}_i(t) := \frac{\text{number of negative instances } \mathbf{x} \text{ in } S \text{ with } x_i > t}{\text{total number of negative instances in } S}. \tag{5}$$

The following shows that this gives a sufficiently accurate estimate for $\beta_i(t)$ if the random sample S is big enough.

Lemma 2.4 *If* $|q-\hat{q}| \leq \frac{\epsilon}{38d}$, $\hat{q} \geq \frac{3\epsilon}{4}$, *and for all* $i = 1,\ldots,d$ *and all* $t \in \mathbf{R}$, $|\alpha_i(t) - \hat{\alpha}_i(t)| \leq \frac{\epsilon p}{36drq}$ *and* $|\gamma_i(t) - \hat{\gamma}_i(t)| \leq \frac{\epsilon p}{36drq}$, *then* $|\beta_i(t) - \hat{\beta}_i(t)| \leq \frac{\epsilon p}{12drq}$.

Proof. We have $|p-\hat{p}| = p\left|1-\frac{\hat{p}}{p}\right| = p\left|1-\left(\frac{\hat{q}}{q}\right)^{1/r}\right|$. If $\hat{q} \geq q$ then $\left(\frac{\hat{q}}{q}\right)^{1/r} \leq \exp\left(\frac{\hat{q}-q}{rq}\right) \leq 1+\frac{\epsilon}{36drq}$ since $\frac{\epsilon}{drq} \leq 1$. If $\hat{q} < q$ then $\left(\frac{\hat{q}}{q}\right)^{1/r} \geq \exp\left(\frac{38}{36}\frac{\hat{q}-q}{rq}\right) \geq 1-\frac{\epsilon}{36drq}$ since $\frac{\epsilon}{drq} \leq 1$. Thus $|p-\hat{p}| \leq \frac{\epsilon p}{36drq}$. Then $|\beta_i(t)-\hat{\beta}_i(t)| \leq |\alpha_i(t)-\hat{\alpha}_i(t)| + |p\gamma_i(t) - p\hat{\gamma}_i(t)| + |p\hat{\gamma}_i(t) - \hat{p}\hat{\gamma}_i(t)| \leq \frac{\epsilon p}{12drq}$. □

Lemma 2.5 ([Vapnik and Chervonenkis, 1971]) *Let* $\mathbf{P}$ *be an arbitrary probability distribution on* $\mathbf{R}$ *and* $f(t) = \mathbf{P}\{x : x > t\}$. *Furthermore define the random variable* $\hat{f}_m(t) = \frac{\#\{1\leq i\leq m : x_i > t, (x_1,\ldots,x_m) \text{ drawn from } \mathbf{P}^m\}}{m}$. *Then for all* $m \geq \frac{32}{\epsilon^2}\log\frac{1}{\delta}$ *the probability that* $|f(t)-\hat{f}(t)| > \epsilon$ *is at most* δ.

Proof. By an adaption of [Vapnik and Chervonenkis, 1971]. □

Lemma 2.6 *If the size of the sample* S *satisfies* $m := |S| \geq 38^2 \cdot 32 \cdot \frac{d^2 r}{\epsilon^2}\log\frac{2d+2}{\delta}$ *and* $\hat{q} \geq \frac{3\epsilon}{4}$ *then the estimates* $\hat{q}$, $\hat{\alpha}_i(t)$, *and* $\hat{\gamma}_i(t)$ *satisfy the conditions given in Lemma 2.4.*

Proof. We use Lemma 2.5. Since $\hat{q}$ is estimated from m independently drawn examples, it follows that $|q-\hat{q}| \leq \frac{\epsilon}{38d}$ with probability $1-\frac{\delta}{2d+2}$. Since $\hat{\alpha}_i(t)$ is estimated from rm independently drawn instances we get $|\alpha_i(t)-\hat{\alpha}_i(t)| \leq \frac{\epsilon}{36dr} \leq \frac{\epsilon p}{36drq}$ with probability $1-\frac{\delta}{2d+2}$ for each $i = 1,\ldots,d$ and all $t \in \mathbf{R}$. Since $\hat{\gamma}_i(t)$ is estimated from the negative instances we first lower bound their number. The probability that from m examples less than $m\left(q-\frac{\epsilon}{38}\right)$ are negative is at most

$\frac{\delta}{2d+2}$. Thus with high probability there are $rmq\left(1-\frac{1}{19}\right)$ negative instances and $|\hat{\gamma}_i(t)-\gamma_i(t)| \leq \frac{\epsilon}{38dr}\sqrt{\frac{19}{18q}} \leq \frac{\epsilon p}{36drq}$ with probability $1-\frac{\delta}{2d+2}$ for each $i=1,\ldots,d$ and all $t \in \mathbf{R}$. Hence with probability δ all estimates are sufficiently accurate. □

2.3 Computational issues, Proof of Theorem 1.6

With all the preceding work the algorithm for calculating a good hypothesis $\hat{C} = \prod_{i=1}^{d}[0,\hat{b}_i]$ can now be described quite easily.

From a sample of size m, $\hat{q}$ and $\hat{p}$ can be calculated by (3) and (4) in time $O(m)$. If $\hat{q} < \frac{3\epsilon}{4}$ the algorithm outputs $\hat{C} := \mathbf{R}^d$. Otherwise it proceeds as follows.

Since the values of $\hat{\alpha}_i(t)$ and $\hat{\alpha}_i(t')$, $t < t'$, differ only if there is an instance $\mathbf{x}$ in the sample with $t < x_i \leq t'$ the values of $\hat{\alpha}_i(t)$ (see (2)) can be calculated incrementally after sorting all instances accordingly to their i-th coordinates. This takes time $O(rm\log(rm))$. The values of $\hat{\gamma}_i(t)$ can be calculated analogously from (5).

Finally, $\hat{b}_i$ can be calculated by considering the i-th coordinates of all instances in ascending order. Then $\hat{b}_i$ is given by the first coordinate x_i which satisfies $\hat{\beta}_i(x_i) \leq \frac{\epsilon\hat{p}}{8dr\hat{q}}$ with $\hat{\beta}_i(\cdot)$ given by (1). This takes time $O(rm)$.

Proof of Theorem 1.6. Obviously the run time of the algorithm is bounded by $O(drm\log(rm)) = O(drm\log m)$. Furthermore, Lemmas 2.1, 2.3, 2.4, and 2.6 show that the calculated hypothesis is ϵ-accurate with probability $1-\delta$. □

3 Conclusion and ongoing research

In this paper we presented an approach to solve the multiple instance learning problem for axis-parallel rectangles. Along these lines similar multiple instance learning problems can be attacked. A particular interesting problem is learning decision trees from multiple instances. Since in general learning decision trees even from single instances is hard, one has to restrict oneself to situations where learning from single instances is possible. It can be shown that in these situations also learning from multiple instances is possible.

Acknowledgments

I want to thank Tom Dietterich and Avrim Blum for very fruitful discussions.

References

[Auer et al., 1997] Auer, P., Long, P. M., and Srinivasan, A. (1997). Approximating hyper-rectangles: Learning and pseudo-random sets. In *Proc. 20th Ann. Symp. Theory of Computing*, pages 314–323. ACM.

[Blum and Kalai, 1997] Blum, A. and Kalai, A. (1997). A note on learning from multiple-instance examples. Unpublished manuscript.

[Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71.

[Long and Tan, 1996] Long, P. and Tan, L. (1996). PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. In *Proceedings of the 1996 Conference on Computational Learning Theory*, pages 228–234.

[Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM*, 27(11):1134–1142.

[Vapnik and Chervonenkis, 1971] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280.

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'