In Proceedings of First Pacific-Asia Conference on Knowledge Discovery and Data Mining.
Pages 252--263. Singapore, February 1997.

EPSILON FOCUSING—A STRATEGY FOR ACTIVE EXAMPLE SELECTION

PARTHA NIYOGI
Artificial Intelligence Laboratory
Massachusetts Institute of Technology, USA

KAH-KAY SUNG
Department of Information Systems & Computer Science
National University of Singapore, Singapore

In most classical formulations of learning from examples, a *passive* learner is presented with examples randomly drawn. Here we discuss the ϵ-focusing strategy to *actively* choose examples for concept learning. We describe the local focused property that functions must have for such a strategy to work. Finally, we develop a heuristic motivated by ϵ-focusing for a practical human face detection task.

1 Introduction

Many problems in machine learning can be cast within a framework where the learner attempts to approximate an unknown target function on the basis of examples provided to it. A formal analysis of such learning problems suggests that the sample complexity of learning is often unreasonably high, i.e., the number of examples the learner would need to collect in order to provably be within an ϵ-width of the target function is much higher than the number of examples typically available in real world learning situations. Crucially, however, the formal analyses typically assume a *passive* learner who is presented examples that are randomly drawn from the space of possible examples. In this paper, we would like to explore the possibility of a more *active* scheme where the learner collects examples that are relevant to the task at hand.

As we shall see, in some cases, knowledge that the function lies within some ϵ-ball (in function space) might directly translate (due to locality properties) into knowledge about the regions of the domain X over which the target function values are uncertain. The learner can then zoom (*epsilon-focus*) in on this region of uncertainty, and sample there. This gives rise to a recursive ϵ-focusing strategy where the learner starts out with a few preliminary examples that provide a loose estimate of the target function. It then derives a region on the domain where it should attempt to collect more examples. These new well chosen examples tighten the learner's estimates and then again allows it to rederive a smaller region of the domain where it should collect examples. The process is repeated in this fashion.

We begin with an illustrative example. We then go on to generalize the approach to arbitrary function classes that share a *local focused* property that allows the epsilon focusing strategy to work. Finally, as a concrete real world application, we describe a face detection system that learns with fewer examples by utilizing such an epsilon focusing strategy in practice.

2 A Simple Illustrative Example

Suppose we want to PAC [7] learn (with (ϵ, δ) accuracy) the following class of indicator functions from $[0,1]$ to $\{0,1\}$.

$$\mathcal{F} = \{1_{[a,1]} : 0 \leq a \leq 1\}$$

Further suppose the distribution P on $[0,1]$ according to which data is drawn is known and is uniform. It is known that a *passive* learner would take at least $\Omega((1/\epsilon)\ln(1/\delta))$ examples to do so. We suggest the following k-step strategy which seeks examples from successively smaller well-focused regions of the domain to learn this class in $\Omega((k/\epsilon^{2/k})\ln(k/\delta)$ examples.

2.1 The ϵ-focusing Algorithm (1)

The learning occurs over k (k can be arbitrarily chosen) stages.

1. Draw enough examples to learn the target with $\epsilon^{1/k}$ accuracy with δ/k confidence. Obtain hypothesis $1_{[\hat{a_1},1]}$.

2. Now ask for examples drawn uniformly at random from the region $[\hat{a_1} - \epsilon^{1/k}, \hat{a_1} + \epsilon^{1/k}]$ and try to learn the target function with $\epsilon^{2/k}$ accuracy with δ/k confidence (with respect to this new distribution over the smaller region). Obtain hypothesis $1_{[\hat{a_2},1]}$.

3. Repeat like step 2, i.e., ask for enough examples drawn uniformly at random from the region $[\hat{a_2} - \epsilon^{2/k}, \hat{a_2} + \epsilon^{2/k}]$ in order to learn the target function to $\epsilon^{3/k}$ accuracy with δ/k confidence. Obtain hypothesis $1_{[\hat{a_3},1]}$. In general at the jth step, ask for examples drawn uniformly at random from the region $[\hat{a_{j-1}} - \epsilon^{(j-1)/k}, \hat{a_{j-1}} + \epsilon^{(j-1)/k}]$ to learn the target to within $\epsilon^{j/k}$ accuracy with δ/k confidence. Obtain hypothesis $1_{[\hat{a_j},1]}$.

4. Stop with hypothesis $1_{[\hat{a_k},1]}$.

2.2 The Spirit of ϵ-focusing

Rather than learn at one step by collecting the requisite number of examples for (ϵ, δ) learning, the method of ϵ-focusing collects a few examples to make a loose ($\epsilon^{1/k}$ accurate) estimate—it then derives a region of interest ($[\hat{a_{j-1}} - \epsilon^{(j-1)/k}, \hat{a_{j-1}} + \epsilon^{(j-1)/k}]$ at the jth step) from where it collects further examples. It keeps refining its estimates by these further examples and eventually narrows down to the target function. For the example considered, the sample complexity is provably reduced to $O((k/\epsilon^{2/k})\ln(k/\delta))$. This is lower than the lower bound on *passive* learning.

There are two useful directions we pursue in the rest of this paper. First, on a theoretical note, we generalize the approach to arbitrary function classes and describe conditions necessary for the provable reduction in sample complexity by ϵ-focusing. Second, on a more practical note, we describe a "boot-strapping" example selection strategy that is motivated by ϵ-focusing to improve the performance of a face detection system.

3 Generalization of Epsilon-focusing

It is of interest to characterize general conditions on function classes for which the ϵ-focusing strategy would yield superior performance. We introduce the notion of localized function classes. These classes which have a *local focusing* property can be learned faster by the method of ϵ-focusing as we show below.

3.1 Localized Function Classes

In the illustrative example provided, the function class had the property that knowledge of the distance between any two functions f and g in $\mathcal{F}$ (in the d_P metric) allowed us to focus in on a region of interest in the domain $X = [0, 1]$ where f and g differ. We formalize this notion to derive a general bound on sample complexity for the ϵ-focusing strategy.

Let $\mathcal{F}$ be a concept class (i.e. class of indicator functions) on some compact domain X. Let P be the uniform distribution on this domain, i.e., the distribution which corresponds to the normalized Lebesgue measure on it. We define the usual $L_1(P)$ distance metric on the space functions by:

$$d_P(f, g) = \int_X |f - g| P(x) dx$$

We define the *local focusing* property of such an arbitrarily defined concept class as follows:

Definition 1 *For a given f belonging to some concept class $\mathcal{F}$ on X, and for any given $\epsilon > 0$, its ϵ-region of interest, $\mathcal{R}_\epsilon(f)$ is given by*

$$\{x \in X | f(x) \neq h(x) \text{ for some } h \in \mathcal{F} \text{ such that } d_P(f, h) \leq \epsilon\}$$

Definition 2 *The concept class $\mathcal{F}$ is said to be locally focused with focusing bound g (g is a real valued function taking values on $[0, 1]$) if for every $\epsilon > 0$,*

$$\sup_{f \in \mathcal{F}} Volume(\mathcal{R}_\epsilon(f)) \leq g(\epsilon)$$

Here, $Volume(s)$ for any set $s \subset X$, is simply the volume[a] of that set. We assume that $Volume(X) = 1$. Thus, locally focused classes are those with bounded ϵ regions of interest into which we can focus in the iterative manner of Algorithm (1).

3.2 The General ϵ-focusing Algorithm (2)

The general algorithm to learn such ϵ-focused classes is as follows:

1. Begin with the entire class $\mathcal{F}$, draw examples according to the uniform distribution P on X, (call this P_1) and attempt to learn the target ($f_t \in \mathcal{F}$) to $\epsilon^{1/k}$ with probability at least $1 - \delta/k$. Obtain hypothesis $\hat{f}_1$. Also obtain the reduced set of candidate target functions (version space),

$$\mathcal{F}_1 = \{f \in \mathcal{F} | d_{P_1}(f, \hat{f}_1) \leq \epsilon^{1/k}\}$$

   Finally, also obtain the ϵ-region of interest: $R_1 = \mathcal{R}_{\epsilon^{1/k}}(\hat{f}_1)$.

2. Draw examples according to a uniform distribution on R_1 (call this distribution P_2) and learn the target to $\epsilon^{2/k}/g(\epsilon^{1/k})$ (according to P_2) with probability greater than $1 - \delta/k$. Now obtain hypothesis $\hat{f}_2 \in \mathcal{F}_1$, the reduced version space:

$$\mathcal{F}_2 = \{f \in \mathcal{F}_1 | d_{P_2}(f, \hat{f}_2) \leq \frac{\epsilon^{2/k}}{g(\epsilon^{1/k})}\},$$

   and the new ϵ-region of interest $R_2 = \mathcal{R}_{\epsilon^{2/k}}(\hat{f}_2)$.

[a] From a more formal perspective, one should really replace $Volume(s)$ by the measure on the set s, i.e., $P(s)$. Clearly, $P(X) = 1$. In our case, we assume that $Volume(X) = 1$. Since P is a uniform distribution, i.e., any point in this set is as likely as any other point, it follows that $P(s)$ is simply $Volume(s)$. We will continue to use this notation, but the reader will easily see that P can be used in general, and in fact, need not even be uniform.

3. Repeat step 2. In general, at the jth step, learn the target to $\frac{\epsilon^{j/k}}{g(\epsilon^{(j-1)/k})}$ (according to distribution P_j), and obtain $\hat{f}_j, \mathcal{F}_j$, and R_j in the obvious way.
4. Stop at the kth step and output hypothesis $\hat{f}_k$.

3.3 Proof of Correctness:

Recall that our eventual goal is to learn the unknown target f_t within ϵ accuracy (according to the distance metric d_P) with probability greater than $1-\delta$.

Consider the first step. The target has been learned to $\epsilon^{1/k}$ accuracy with high confidence. The learner's hypothesis is $\hat{f}_1$. Clearly, with high probability (greater that $1-\delta/k$), the target lies within in an $\epsilon^{1/k}$ ball around $\hat{f}_1$ (this is denoted by $\mathcal{F}_1$). According to our definition, all functions in $\mathcal{F}_1$ agree on the region outside of R_1. So we only need to sample the region R_1 which is what we do in the second step.

In the second step, we learn the target to $\epsilon^{2/k}/g(\epsilon^{1/k})$. This is according to a distribution P_2 (uniform on the region R_1). Again, the target, is within an $\epsilon^{2/k}/g(\epsilon^{1/k})$ ball of the hypothesis at this stage ($\hat{f}_2$). Thus,

$$d_{P_2}(\hat{f}_2, f_t) = \frac{Volume(\{x \in R_1 | \hat{f}_2(x) \neq f_t(x)\})}{Volume(R_1)} \leq \epsilon^{2/k}/g(\epsilon^{1/k})$$

But, $Volume(R_1) \leq g(\epsilon^{1/k})$. Therefore:

$$Volume(\{x \in R_1 | \hat{f}_2(x) \neq f_t(x)\}) \leq \epsilon^{2/k}$$

Clearly,

$$d_P(\hat{f}_2, f_t) = Volume(X \setminus R_1)(0) + Volume(\{x \in R_1 | \hat{f}_2(x) \neq f_t(x)\}) \leq \epsilon^{2/k}$$

Thus, after the second step, we see that the target f_t is within $\epsilon^{2/k}$ accuracy (with respect to our original distribution P). By our definition of the local focusing property, we know that $f_t \in \mathcal{F}_2$, and the points on which f_t and $\hat{f}_2$ disagree must lie within R_2.

In general, before the jth step, the points on which the target and the $(j-1)$th hypothesis disagree must lie within R_{j-1}. Since, we sample according to a uniform distribution on this (P_j), and attempt to learn the target to an accuracy of $\epsilon^{j/k}/g(\epsilon^{(j-1)/k})$, by a similar argument,

$$d_{P_j}(\hat{f}_2, f_t) = \frac{Volume(\{x \in R_{j-1} | \hat{f}_j(x) \neq f_t(x)\})}{Volume(R_{j-1})} \leq \epsilon^{j/k} / g(\epsilon^{(j-1)/k})$$

But, $Volume(R_{j-1}) \leq g(\epsilon^{(j-1)/k})$. Therefore,

$$Volume(\{x \in R_{j-1} | \hat{f}_j(x) \neq f_t(x)\}) \leq \epsilon^{j/k}$$

and

$$d_P(\hat{f}_j, f_t) = Volume(X \setminus R_{j-1})(0) + Volume(\{x \in R_{j-1} | \hat{f}_j(x) \neq f_t(x)\}) \leq \epsilon^{j/k}$$

Thus, after the jth step, the learner has learned the target to $\epsilon^{j/k}$ accuracy. Further, according to our definition of the local focusing property, the points on which the learner and target disagree must lie within the set $R_j = \mathcal{R}_{\epsilon^{j/k}}(\hat{f}_j)$.

Clearly, after the kth step, the learner will have learned the target to ϵ accuracy. The only way in which this would not happen is if the learner failed made a mistake in any one of the steps. The probability of making a mistake in each step is δ/k. The probability of making a mistake in any one is bounded by δ. Thus, the learner would have identified the target to ϵ accuracy with confidence greater than $1 - \delta$. □

3.4 Sample Complexity

By the standard Vapnik Chervonenkis theorem [8] we see that at the jth stage, the learner will have to draw at most $O((g^2(\epsilon^{(j-1)/k})/\epsilon^{2j/k}) \ln(k/\delta))$ examples to satisfy the learnability requirement of that stage. The total number of examples the learner needs would be

$$O\left(\sum_{j=1}^{k} \frac{g^2(\epsilon^{(j-1)/k})}{\epsilon^{2j/k}} \ln(k/\delta)\right)$$

3.5 Remarks

Now we are in a position to re-evaluate our simple example from this general perspective. It is easy to see that

1. **Opening Example:** For an arbitrary $f_a = 1_{[a,1]}$, we see that $\mathcal{R}_\epsilon(f_a) = [a - \epsilon, a + \epsilon]$. Clearly, $g(\epsilon) = 2\epsilon$. The corresponding sample complexity is $O((k/\epsilon^{2/k}) \ln(k/\delta))$.

2. **Box Functions:** Consider the following class of indicator functions on $[0,1] : \mathcal{F} = \{1_{[a,b]} : 0 \leq a \leq b \leq 1\}$. For an arbitrary $f_{a,b} = 1_{[a,b]}$, we see that $\mathcal{R}_\epsilon(f_{a,b}) = [a-\epsilon, a+\epsilon] \cup [b-\epsilon, b+\epsilon]$. Clearly, $g(\epsilon) = 4\epsilon$. The sample complexity $O((k/\epsilon^{2/k})\ln(k/\delta))$ follows.

3. **Extensions to n-dimensions:** Suppose one considers the PAC learning of a concept class of hyperplanes dividing an n-simplex into two regions. Essentially, the hyperplane cuts the simplex at its edges. Consequently, along each edge, the points on one side of the cut are labelled 0, while the points on the other side are labelled 1. Such a problem can then be reduced to n one-dimensional problems and a provable sample complexity of $O((n^2/\epsilon^{1/k})\ln(kn/\delta))$ In view of the fact that we have an effective ϵ-focusing strategy for box functions, we can even address concept classes represented by multilayer perceptrons with two hidden layers.

4. **Handling misclassification noise:** The ϵ-focusing strategy in this part has been developed for a noise-free case. Extensions to cover a situation with a bound on the misclassification noise (the label of the example can be flipped with probability at most η) can be considered as well. In this case, the ϵ-region of interest can be bounded in terms of the misclassification noise η.

Finally, some natural questions arise at this stage. First, given that locally focused classes can be efficiently learned by ϵ-focusing, we need to know what kinds of concept classes have the locally focusing property in the first place? Second, given the existence of the locally focusing property, how easy is it to compute the ϵ-region of interest $\mathcal{R}_\epsilon$ for such concept classes. Further research on these questions is awaited.

4 A Real Application — Learning Human Face Detection

We now describe a real application of the *epsilon-focusing* idea on training a classifier for human face detection [5] [6] Given a digitized image, we want an algorithm that returns the location and size of each human face in the image. For now, we shall just limit ourselves to vertical frontal views of human faces.

Computational efficiency aside, one can reduce the face detection problem to searching an image exhaustively for all square patches having the general structure of Figure 1(a). Our task then, is to train a classifier to recognize such patterns using a well chosen database of positive (face) and negative (non-face) examples. Ideally we would like to train the system with as large a set of "face" and "non-face" examples as possible, in order to attain classification

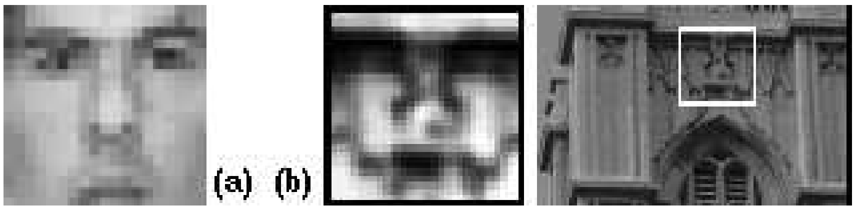

Figure 1: **(a):** A "canonical" face pattern. **(b):** An example of a non-face pattern that resembles a human face when viewed in isolation. The *epsilon-focusing* example selection strategy seeks out these patterns with high information value.

with minimal error. Unfortunately, we must constrain the size of our training database because of limited disk space and training time.

How does one build a comprehensive but tractable database of "face" and "non-face"" patterns? For "face" patterns, we simply collect all the frontal views of faces we can find in mugshot databases and other image sources. Because we do not have access to many mugshot databases, the number of "face" examples remains manageably small. For "non-face" patterns, the task is more tricky. In essence, every square non-face window pattern of any size in any image is a valid "non-face" example. Clearly, even with a few source images, our set of "non-face" patterns can grow intractably large if we are to include all valid "non-face" patterns in our training database.

We borrow a few ideas from ϵ-focusing to derive a heuristic training strategy for the problem described above. First, in the spirit of ϵ-focusing, we will not attempt to train the classifier in one step—rather, we break the the training process into several stages. At each stage, if we were to follow ϵ-focusing *exactly* then, we should collect examples from the ϵ region of interest, i.e., $R_j = \mathcal{R}_{\epsilon^{j/k}}(\hat{f}_j)$. Note that this set is essentially the set of points on the domain that are differently classified by some function that is "close" to the current hypothesis $\hat{f}_j$. Points that are classified similarly are of no consequence in the training process and carry no information value. Therefore, we concentrate only on the misclassified examples to retrain our system. Although, we should ideally consider only those misclassified examples that are close to the boundary, in practice we actually consider all misclassified examples. This is because we do not have a formal bound yet on the ϵ-region of interest for our hypothesis class (a class of feedforward RBF/MLP hybrid networks).

This gives rise to the following strategy:

1. Start with a small and possibly incomplete set of "non-face" examples

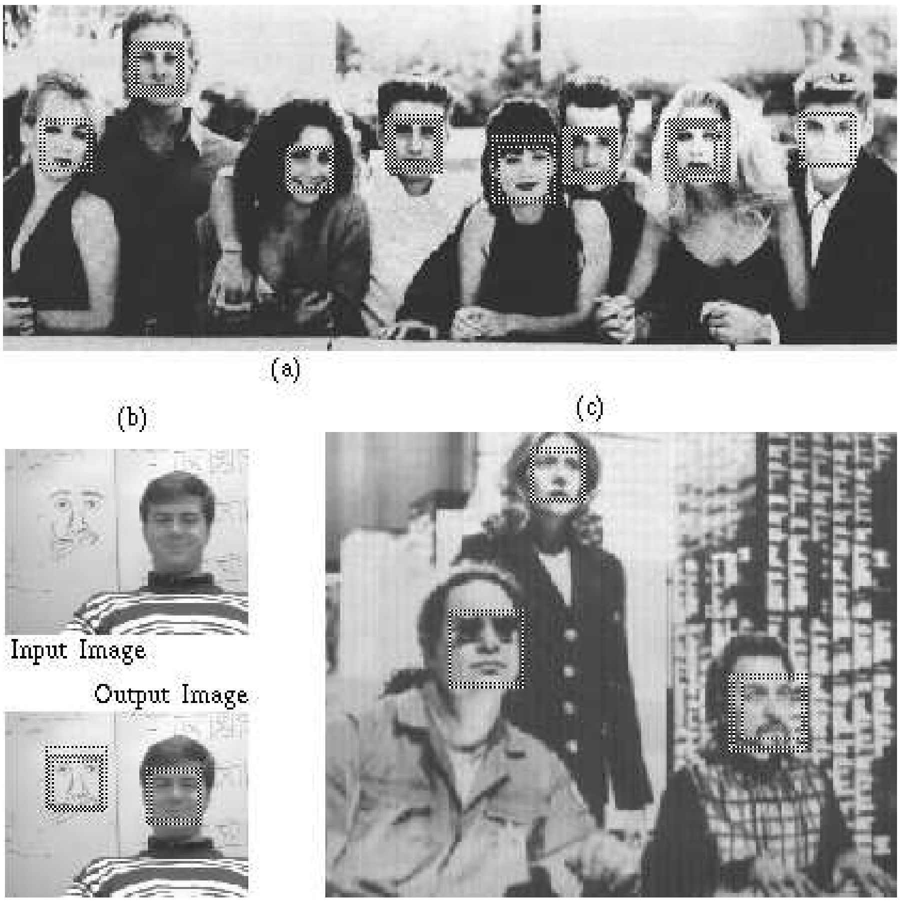

Figure 2: Some face detection results by the (Sung and Poggio 94) frontal view system, trained using a strategy similar to ϵ-focusing. **(a):** An image with many (8) frontal faces. The system finds all 8 faces successfully. **(b):** An input-output image pair. The same system detects real human faces and hand drawn faces. **(c):** The system finds faces successfully over multiple scales. There are no false detects even in a background with complex texture.

Figure 3: Face detection results on two images. **Left:** The system trained without ϵ-focusing produces several false alarms. **Right:** The ϵ-focusing system is trained with a better set of "non-face" examples and does not produce any false alarms on both images. Notice that the faces in both output images are enclosed by light dotted-squares, indicating successful detection.

in the training database. Train the neural net classifier with the current database of examples.

2. Run the face detector on a sequence of images with no faces. Collect all the patterns that the current system wrongly classifies as "faces" (see for example Figure 1(b)). Add these mis-classified patterns to the training database as new negative examples. Retrain to obtain updated classification system.

3. Return to Step 2.

We trained one such system in two iterations ($k = 2$). The final database consists of 4150 face patterns and about 40,000 non-face patterns, of which 6000 were selected in the second iteration. On a test set of 23 cluttered images with 147 faces, the system missed 23 faces and produced 13 false alarms. As a control, we trained a second system (without ϵ-focusing) using the same 4150 face patterns and another 40,000 randomly selected non-face patterns. This yielded 15 misses and 44 false alarms.

First note that system trained with ϵ-focusing has a lower total error rate than the control trained without ϵ-focusing. It has a slightly higher miss rate for faces but a much lower false alarm rate for non-faces. This is because we have used ϵ-focusing only to select better "non-face" examples while leaving the "face" examples unchanged. Consequently we are better able to reject non-faces at a slight expense of correctly detecting faces.

5 Conclusions

We have described a general strategy for selecting examples that provably reduces the sample complexity of learning for functions that have a local focused property. We proceeded to develop a heuristic for a real world human face detection problem with superior empirical performance.

References

1. P. Niyogi. The Informational Complexity of Learning from Examples. Ph.D. Thesis, Massachusetts Institute of Technology. 1995.
2. K. Sung and P. Niyogi. Active Learning for Function Approximation. In *Proc. Advances in Neural Information Processing Systems 7*. April 1995.
3. K. Sung and P. Niyogi. Active Learning the Weights of a RBF Network. In *Proc. Neural Networks for Signal Processing*. September 1995.

4. K. Sung and P. Niyogi. A Formulation for Active Learning with Applications to Object Detection. Tech Report AIM-1438, Artificial Intelligence Lab, Massachusetts Institute of Technology, May 1996.
5. K. Sung and T. Poggio. Example-based Learning for Human Face Detection. Tech Report AIM-1521, Artificial Intelligence Lab, Massachusetts Institute of Technology, December 1994.
6. K. Sung and T. Poggio. Learning Human Face Detection in Cluttered Scenes. In *Proc. Intl. Conf. on Computer Analysis of Images and Patterns*. September 1995.
7. L. Valiant. A Theory of the Learnable. In *Proc. of the 1984 STOC*. 1984.
8. V. Vapnik. Estimation of Dependencies Based on Empirical Data. Springer-Verlag, Berlin. 1982.