ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/228671734

Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper?

Article · January 2005

CITATIONS
15

READS
784

2 authors, including:

Christophe Giraud-Carrier
Brigham Young University - Provo Main Campus
271 PUBLICATIONS **3,246** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Adolescent Text Messaging Study View project

PhD Thesis View project

All content following this page was uploaded by Christophe Giraud-Carrier on 23 May 2014.

The user has requested enhancement of the downloaded file.

Toward a Justification of Meta-learning: Is the No Free Lunch Theorem a Show-stopper?

Christophe Giraud-Carrier CGC@CS.BYU.EDU
Department of Computer Science, Brigham Young University, Provo, UT 84602

Foster Provost FPROVOST@STERN.NYU.EDU
Stern School of Business, New York University, New York, NY 10012-1126

Abstract

We present a preliminary analysis of the fundamental viability of meta-learning, revisiting the No Free Lunch (NFL) theorem. The analysis shows that given some simple and very basic assumptions, the NFL theorem is of little relevance to research in Machine Learning. We augment the basic NFL framework to illustrate that the notion of an Ultimate Learning Algorithm is well defined. We show that, although cross-validation still is not a viable way to construct general-purpose learning algorithms, meta-learning offers a natural alternative. We still have to pay for our lunch, but the cost is reasonable: the necessary fundamental assumptions are ones we all make anyway.

1. Introduction

Originally introduced to the Neural Network Community, the No Free Lunch (NFL) theorem (Wolpert & Macready, 1995; Wolpert, 2001) was contextualized and brought to the attention of the Machine Learning community in the form of a Law of Conservation for Generalization Performance (LCG): When taken across all learning tasks, the generalization performance of any learner sums to 0 (Schaffer, 1994). The dramatic presentation of the LCG at the 1994 *International Conference on Machine Learning* provoked strong reactions, and an email discussion subsequently was launched by Pazzani in the *Machine Learning List* to engage the community in thoughtful consideration and feedback. That discussion went on fairly consistently for about three months, involving many prominent Machine Learning researchers of the time (see (ML-List, 1994), Numbers 19-27). Although much insight arises from this thread of semi-formal email exchanges, we know of no attempt to distill it into a coherent whole, readily accessible to the community at large. With time, the NFL theorem has almost become fossilized. It is cited for different purposes, but it often seems to be poorly understood.

As some researchers have begun to explore meta-learning as a means of designing robust learning systems, others have been quick to point to the NFL theorem as the sure show-stopper. We revisit the NFL theorem, building on the *Machine Learning List*'s discussion. We show that given some simple and very fundamental assumptions, it is of little relevance to research in Machine Learning.

We make explicit these assumptions, which underlie Machine Learning research. We show that, although cross-validation still is not a viable way to a construct "general-purpose" learning algorithm,[1] meta-learning offers a natural alternative. Moreover, we argue that, to be consistent as machine learning researchers, we should prefer meta-learning over manual construction of general-purpose learning algorithms. Lastly, we discuss how the necessary fundamental assumptions are assumptions we all make anyway.

2. NFL Revisited

As a simple illustration of the NFL theorem, consider the simple space, $\mathcal{F}$, of binary functions defined over $\mathbb{B}^3 = \{0,1\}^3$, and assume that the instances of set $Tr = \{000, 001, \ldots, 101\}$ are observed, whilst the instances of set $Te = \mathbb{B}^3 - Tr = \{110, 111\}$ constitute the

Appearing in *Proceedings of the ICML-2005 Workshop on Meta-learning*, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).

[1] An algorithm designed to apply beyond a specific task or tasks, but without a specific, non-trivial characterization of the class(es) of task to which it should be applied.

	Inputs			f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	...
Training Set	0	0	0	0	0	0	0	0	0	0	0	0	0	...
	0	0	1	0	0	0	0	0	0	0	0	0	0	...
	0	1	0	0	0	0	0	0	0	0	0	0	0	...
	0	1	1	0	0	0	0	0	0	0	0	0	0	...
	1	0	0	0	0	0	0	0	0	0	0	1	1	...
	1	0	1	0	0	0	0	1	1	1	1	0	0	...
Test Set	1	1	0	0	0	1	1	0	0	1	1	0	0	...
	1	1	1	0	1	0	1	0	1	0	1	0	1	...

Figure 1. Sample Train/Test Setting for Binary Functions over 3 Boolean Variables

off-training set (OTS) test set, as depicted in Figure 1. The NFL theorem, or LCG, in this setting shows that, averaged across all functions, $f_1, f_2, \ldots, f_{256} \in \mathcal{F}$ (i.e., all labelings of instances), the behavior on Te of any learner trained on Tr is that of a random guesser.[2]

A quick examination of Figure 1 makes this result intuitive and obvious. Consider functions f_1 through f_4 in Figure 1. For all 4 functions, Tr is the same. Hence, provided any deterministic learner L, the model induced by L from Tr is the same in all 4 cases. It follows immediately that, since the associated Te's span all possible labelings of the OTS instances, for any OTS instance any model will be correct for half the functions and incorrect for the other half. Indeed, every classifier will have accuracies of 100%, 50%, 50%, and 0% across the four functions. The argument is easily repeated across all such subsets of 4 functions, giving the overall result.[3]

It then becomes apparent that the NFL theorem in essence simply restates Hume's famous conclusion about induction having no rational basis:

> there can be no *demonstrative* arguments to prove, *that those instances, of which we have had no experience, resemble those, of which we have had experience*....Thus not only our reason fails us in the discovery of the *ultimate connexion* of causes and effects, but even after experience has inform'd us of their *constant conjunction*, 'tis impossible for us to satisfy ourselves by our reason, why we shou'd extend that experience beyond those particular instances, which have fallen under our observation. We suppose, but are never able to prove, that there must be a resemblance betwixt those objects, of which we have had experience, and those which lie beyond the reach of our discovery. (Hume, 1740)

All other things being equal, given that all one has seen is Tr and its labeling, there is no rational reason to prefer one labeling of Te over another.

At this point, we bring up an important misconception about the NFL theorem, which although previously addressed continues to prevail in the community. The above line of argument seems to hang on the assumption that all functions are equally likely, i.e., uniformly distributed. Technically, this is not true. Schaffer's LCG formulation involves a summation, which says nothing about the distribution of functions. As pointed out by Gordon and Spears (ML-List, 1994), although the result clearly holds under the uniform distribution, there are other non-uniform distributions that satisfy the LCG.[4] The crucial and most powerful contribution of the NFL theorem is pointing out that whenever a learning algorithm performs well on some function, as measured by OTS generalization, it must perform poorly on some other(s).

In his justification for the use of OTS generalization in the NFL theorem, Wolpert claims, among other things, that OTS generalization is the truly interesting measure, i.e., what we really want to know is how

[2] Note here that generalization performance consists of two components: expected performance over instances that have already been encountered and expected performance over instances that have not yet been encountered (i.e., OTS). The former is not trivial, since previously encountered instances are not guaranteed to have the same label as in the training set—so there still is an important statistical estimation task. However, we restrict attention here to the OTS performance as that is the focus of the NFL theorem.

[3] A similar intuitive argument is presented in (Duda et al., 2001).

[4] Wolpert's (Wolpert, 2001) formulation of the NFL theorem for supervised learning does specify uniform averaging over all functions. The NFL theorem for general search has been "sharpened," specifying necessary and sufficient conditions for its applicability (Schumacher et al., 2001; Igel & Toussaint, 2004).

well our algorithm performs beyond what it has seen (Wolpert, 2001). Practically speaking, even granting that it is worthwhile to separate OTS cases from cases previously encountered, *expected* OTS generalization is the true quantity of interest—emphasizing that some OTS cases may be more likely to be encountered than others. Following this same line of argument at the meta-level, we claim that the real interesting measure at the meta-level is *expected* generalization performance, i.e., how well our algorithm is likely to perform on functions beyond those it has already experienced, given the particular distribution from which the functions we will encounter are drawn. As suggested by Jenkins and aptly demonstrated by Holte (ML-List, 1994), the NFL theorem says nothing about such expected generalization performance. Hence, although interesting in its own right, the NFL theorem is in this sense irrelevant to Machine Learning research.

3. The Basic Assumption(s) of Machine Learning

It is clear that effective general-purpose learning systems exist. Daily human experience provides irrefutable evidence that humans can learn. As Hunter states (ML-List, 1994):

> ...human performance [is] a clear existence proof that it is possible to exhibit useful generalization performance in the extremely broad class of complex and difficult learning problems that *tend to appear in our world.* (emphasis added).

This argument is implicit in the basic assumption of inductive learning, which although widely accepted, is rarely, if ever, explicitly stated.

Definition 1. *The* weak assumption *of Machine Learning is that the process that presents us with learning problems, call it Ω, induces a non-uniform probability distribution, p^{Ω}, over the f_i's.*[5]

In other words, Machine Learning researchers and practitioners, like statisticians, do not apply the *principle of indifference* (Kneale, 1949), but instead assume that some functions are, in reality, more likely than others. Importantly, *this weak assumption is sufficient to claim that there exist some algorithms that are better than others* (as suggested by Jenkins (ML-List, 1994)). We are unaware of anyone ever suggesting that the weak assumption does not hold, yet the NFL theorem often is cited as proving that there can exist no general-purpose learning algorithms. There is a subtle-but-important difference between whether such an algorithm can exist, and whether we would know it if we were to see it.

Researchers working on the design of general-purpose algorithms make an even stronger assumption about the world.

Definition 2. *The* strong assumption *of Machine Learning is that p^{Ω} is explicitly or implicitly known, at least to a useful approximation.*

Indeed, each learning algorithm contains a bias, which embodies this strong assumption. In some sense, the assumed p^{Ω} corresponds to the area of expertise of the learning algorithm (Bensusan & Giraud-Carrier, 2000).

Having established the above, we now turn to the definition of what one might call an Ultimate Learning Algorithm (ULA). First, we note that:

Lemma 1. *Knowing $p(f)$, the probability of encountering an arbitary function f, is equivalent to knowing $p(c|e)$, the probability of class membership for an arbitrary example e.*

Proof. *Let n be the size of the input space and m the size of the function space. Let c be the class membership of a given example e. By definition, for any function f_k:*

$$\begin{array}{rcl} p(f_k) & = & p(c = f_k(e_1)|e_1, \ldots, c = f_k(e_n)|e_n) \\ & = & \Pi_{i=1}^{n} p(c = f_k(e_i)|e_i) \\ & = & \Pi_e p(c = f_k(e)|e) \end{array}$$

Hence, if $p(c|e)$ is known for all e, then clearly, so is $p(f_k)$, and more generally $p(f)$. Similarly, for any e_i, c_j:

$$\begin{array}{rcl} p(c_j|e_i) & = & p(f_1(e_i) = c_j)p(f_1) + \ldots + p(f_m(e_i) = c_j)p(f_m) \\ & = & \sum_{k=1}^{m} p(f_k(e_i) = c_j)p(f_k) \\ & = & \sum_{f_k : f_k(e_i) = c_j} p(f_k) \end{array}$$

Hence, if $p(f)$ is known for all f, then clearly, so is $p(c_j|e_i)$, and subsequently $p(c|e)$. □

Given a training set, a learning algorithm, L, induces a model, M, which defines a class probability distribution, p, over the instance space.

Definition 3. *An Ultimate Learning Algorithm, ULA, is a learning algorithm that induces a model $M^{\star}$, such that:*

$$\forall M' \neq M^{\star} \;\; E(\delta(p^{\star}, p^{\Omega})) \leq E(\delta(p', p^{\Omega}))$$

where the expectation is computed for a given training/test set partition of the instance space, over the

[5] To be exact, the definition should have the stronger requirement that p^{Ω} does not satisfy the LCG. We ignore this subtlety here.

entire function space, and delta is some appropriate distance measure.

Finding a ULA thus consists of finding a learning algorithm whose induced models closely match our world's underlying distribution of functions. In the context of the LCG and ultimate learning algorithms, we thus concur with Hartley (ML-List, 1994) who was quick to note that there are two relevant definitions of the word "universal"[6]: (1) applicable independent of any assumptions, and (2) applicable throughout the entire universe. "The first is what most mathematicians mean by universal, and by this definition the conservation law rules out any useful generalization. However when asking about the real world it is the second definition that is important. What could happen in other conceivable universes is of no possible interest to us."[7]

We note however that the definition of ULA is not in complete contradiction with the NFL theorem. Indeed, in some kind of asymptotic way, ULA = Ω, so that if Ω is such that p^{Ω} is in fact uniform, then ULA is a random guesser, as expected. We now turn to the question of how to build a ULA.

4. How to Build an Ultimate Learning Algorithm

In seeking to design a ULA, researchers have used two approaches: cross-validation and manual algorithm design. We show that the first is not viable. We suggest that the second is a possibility, but argue that it makes stronger assumptions than we might like, and also is somewhat at odds with the philosophy of machine learning. We propose meta-learning as a viable alternative.[8]

4.1. Cross-validation Model Selection

Cross-validation is regularly used as a mechanism to select among competing learning algorithms, as illustrated in Figure 2. As pointed out by Wolpert (Wolpert, 2001), however, cross-validation is also subject to the NFL theorem. This is easily seen again from Figure 1. Since Tr does not change over f_1 through f_4, cross-validation always selects the same best learner in each case and the original NFL theorem applies. Here, we show further that a No Free Lunch result holds for cross-validation, even under the weak assumption of Machine Learning.

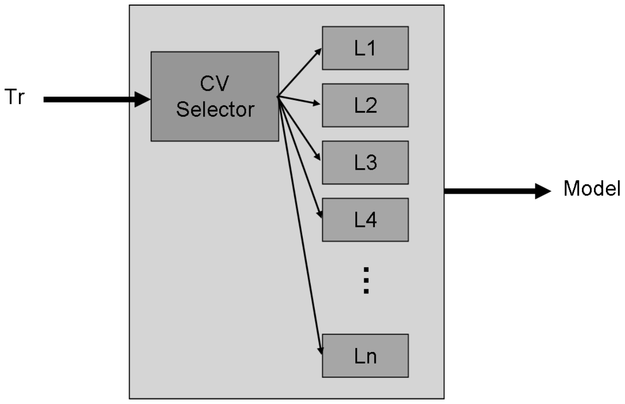

Figure 2. CV Meta-selector

Theorem 1. $\sum_i EGP_{CV}(f_i) = 0$

where EGP is the expected OTS generalization performance.

Proof. *For simplicity, consider the setting of Figure 1. Let $\mathcal{L}$ be an unbiased set of deterministic learning algorithms that a selector can choose from via cross-validation. Given a training set, each $l \in \mathcal{L}$ generates only one $f \in \mathcal{F}$. Since $\mathcal{L}$ is unbiased, it contains an equal number of learners for each possible $f \in \mathcal{F}$. Consider any subset $F \subset \mathcal{F}$ of functions with identical training set, T. Clearly, cross-validation has no basis in T, other than random fluctuations, to choose one $f \in F$ over any other. Since $\mathcal{L}$ is unbiased, cross-validation will pick each with equal probability, regardless of the prior distribution of $f \in \mathcal{F}$. It follows that the expected OTS generalization performance of the CV meta-selector over F is 0. Repeating the argument over all such subsets F gives the result.* □

It follows that cross-validation cannot generalize and thus can not be used as a viable way of building an ultimate learning algorithm.

4.2. Manual Algorithm Design

The traditional approach in Machine Learning research is to design one's own algorithm. This is generally motivated by the presence of one or more tasks

[6] Our choice of "ultimate" deliberately avoids this ambiguity of the word "universal" and makes clear the connection of a ULA to the weak assumption.

[7] Hartley goes even further to show that, given the size (in terms of atoms) of the universe as we know it, it is simply impossible to represent all possible functions of even only about 200 attributes, so that the utility of the NFL theorem is questionable. "If there exists an algorithm for which the sum is sufficiently different from zero this would be a universal generalization algorithm in any useful sense of the phrase."

[8] It remains rather surprising to us that throughout the extended discussions of NFL theorem by machine-learning researchers, a process of meta-learning does not seem ever to have been mentioned.

that one wishes to learn, and for which no existing algorithm seems to perform to the desired level. The researcher then sets out to design—generally with much (unreported) trial-and-error—a new learning algorithm that performs well on the target tasks and others that are available for evaluation. This approach depends upon the strong assumption of machine learning, that p^{Ω} is known well enough to be incorporated by the algorithm's inductive bias.

This expertise-driven, knowledge-based strategy results in implicit meta-learning (see below) by the research community and is an essential part of the science of Machine Learning. Historically, new algorithms tend to be designed specifically to overcome limitations of known algorithms as they are discovered (e.g., new tasks arise for which no existing algorithm seems to be suitable). In the process, the community's knowledge about learning increases.

4.3. Meta-learning

In the last decade, some researchers have attempted to design better learning algorithms by applying "meta-learning"—learning for model selection at the meta-level (e.g., see (Rendell & Cho, 1990; Michie et al., 1994; Pfahringer et al., 2000; van Someren, 2001; Vilalta et al., 2004). As stated in (Vilalta et al., 2004), "meta-learning differs from base-learning in the scope of the level of adaptation; whereas learning at the base-level is focused on accumulating experience on a specific learning task (e.g., credit rating, medical diagnosis, mine-rock discrimination, fraud detection, etc.), learning at the meta-level is concerned with accumulating experience on the performance of multiple applications of a learning system.... Briefly stated, the field of meta-learning is focused on the relation between tasks or domains and learning strategies."

Meta-learning, in the context of model selection, consists of applying learning mechanisms to the problem of mapping classification tasks to algorithms. Let L be a set of learning algorithms for classification and T be a set of classification tasks such that for each $t \in T$, $b_L(t)$ represents the algorithm in L that performs best on t and $c(t)$ denote the characterization of t by some fixed mechanism. Then, meta-learning takes the set $\{< c(t), b_L(t) >: t \in T\}$ as a training set and induces a meta-model that, for each new classification task, predicts the model from L that will perform best.

In our framework, meta-learning is cast as the task of learning $\hat{p}^{\Omega}$. It assumes that it is possible to gather training data at the meta-level to learn $\hat{p}^{\Omega}$, and use that information to select among base-level learners, as depicted in Figure 3.

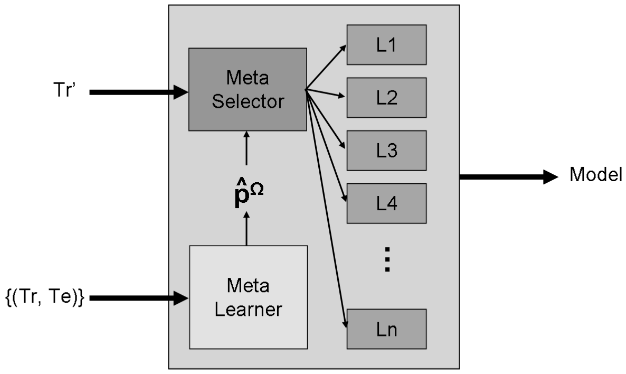

Figure 3. Meta-learning Selector

Importantly, meta-learning assumes only a bias for learning $\hat{p}^{\Omega}$. There are important potential implications of this view. For example, once $\hat{p}^{\Omega}$ is learned, one may wonder whether base-level learning can add any value. Should the task of meta-learning be simply to learn $\hat{p}^{\Omega}$ (rather than a mapping from tasks to learners)? We revisit this and other implications for research in the penultimate section. Before that, let us discuss whether the assumptions we have to make are reasonable.

5. Assumptions and Limitations

We began by asserting that the NFL theorems are in an important sense mostly irrelevant to machine learning. Researchers designing or applying learning algorithms intended to be useful across many tasks must believe one (or both) of two things: (1) There is something about a learning task, separate from the available training data, that allows the selection of one learning algorithm over another, or (2) one of the basic assumptions (described above) of machine learning holds. Clearly, practicing researchers believe both. We believe that for many problems, domain knowledge allows the design of a representation, including important features, such that the "general-purpose" learning algorithms that have appeared to work well in the past again will work well.

Researchers (and practitioners) also have general beliefs about Ω. In particular, we believe that our learning algorithms are unlikely to be faced with bizarre functions. "Bizarre" could be defined by begging the question: those functions for which the algorithms are unlikely to work. So, we generally believe that it is highly unlikely that linear regression will be just as good as it is bad on practical problems with which we will be faced. Beyond a simple circular argument, this belief may be a confidence in a practitioner's intuition about the problems to choose. Domain knowledge may restrict attempts to learn random, chaotic, or otherwise bizarre functions.

Nonetheless, we also should be clear that even with meta-learning our lunch is not free. NFL results apply at the meta-level just as they apply at the level of particular learning tasks. We should consider carefully what this means. Meta-learning departs from the setting of the NFL theorems, in that it takes into account prior observations from p^{Ω}. However, although it is beyond the scope of this workshop paper, it is reasonable to conjecture that there is a direct analog of the basic NFL theorem to the meta-level—showing that all meta-learners have equivalent performance given some (probably debatable) averaging across Ω's.

We claim that the assumptions that must be made for meta-learning are considerably more natural than those that must be made for manual algorithm design. In addition to the partially circular notions above, we all also hold deep-rooted, intuitive notions of bizarre functions. For example, (among many things) we believe that variables that never have exhibited any relevance are more likely than not to continue to be irrelevant. After many years of experience, we would find it bizarre for the date to play a role in whether crows are black or whether gravity will be attractive or repulsive.[9] Harkening back to Hume, there is no rational reason for these beliefs, which of course is the "riddle" of induction. However, implicit in Western thinking is that if we were to make only one assumption, it would have to be that induction is valid—that we can generalize from what we have seen to things we have yet to encounter. This is the fundamental assumption of science as we practice it. It also is fundamental to our being able to live at ease in the world, not constantly worrying for example that the next time we step on a bridge it will not support our weight.

Moreover, we believe that this line of argument is in sharp contrast to current views of the import of the NFL theorems. The machine learning textbook by Duda et al. (Duda et al., 2001) is exceptional (unique?) in its inclusion of the NFL theorems. However, the position we take here is at odds with the conclusions they draw. For example, they say, "if we make no prior assumptions about the nature of the *classification task*, can we expect any classification method to be superior or inferior overall? ... As summarized in the No Free Lunch Theorem, the answer to (this) and several related questions is 'no"' (emphasis added). And later, "This ... stresses that it is the *assumptions* about the learning domains that are relevant" (emphasis in the original). Our stance is that assumptions (or domain knowledge) about the nature of the classification task in fact are not necessary to yield good performance (although they may be quite helpful). In fact, we assert that assumptions about $p(f)$ are not even necessary.

So how do we pay for our lunch? We do need to make hyper-assumptions about $p(f)$, to enable meta-learning. This corresponds to an inductive bias at the meta-level. For example, we might have a hyper-bias that says: prefer p(f)'s that give higher probability to simpler functions; or, returning to the gravity example, prefer p(f)'s that give higher probability to concepts that do not change abruptly over time.[10]

There seems to be ground in nature to believe that such hyper-assumptions are reasonable. Building on Goodman's notion of predicate entrenchment, Russell identified what he calls *high-level regularities* in nature, i.e., meta-rules that lead us to perform induction at the base level *this* way rather than *that* way (Russell, 1986). Hence, there is a bias favoring some $p(f)$'s rather than others.

6. Extensions and Research Implications

Various extensions and implications for meta-learning research follow from this preliminary analysis. We mention several here.

- Meta-learning's assumption of a bias for learning p^{Ω} clearly falls between the Weak Assumption and the Strong Assumption of Machine Learning. We need to introduce a Not-so-weak Assumption, that is in line with the discussion in the previous section.
- Our purpose in learning a classifier is to produce as good an estimation of $p(c|e)$ as possible. Consider Lemma 1 again. It says that knowing $p(c|e)$ and $p^{\Omega}(f)$ are equivalent. Therefore, we should carefully consider why meta-learning for model selection is done as it is currently done: learning to select a base-level learning algorithm that then will be run on the training data to induce a model, that then will be applied to the test data.

  Is there value in doing the base-level learning? Perhaps we should prove Lemma 2: *If we know*

[9] Compare Goodman's "grue" paradox and related discussions (Goodman, 1946; Goodman, 1983); one might attribute the origin of the Y2K problem in part to the unnaturalness of such a concept.

[10] This is why we all favor green over grue in Goodman's analysis of Hume's riddle of induction.

$p^{\Omega}(f)$, *Tr only rules out certain f's. By Lemma 1, inference is just MAP classification based on* $p^{\Omega}(f)$, *yielding* $p(c|e)$ *over the remaining f's.* By knowing $p^{\Omega}(f)$, we know all the $p(f)$'s (Lemma 1). We claim that seeing the training set will not change the relative probabilities of the remaining f's. We then might follow up with a Theorem: *Once we know* $p^{\Omega}(f)$*, base-level "learning" adds no value.*

And a corollary to the above is: *In principle, there is no need for meta-learning to tell us which algorithm to use. Meta-learning should just learn* $p^{\Omega}(f)$ *and apply it directly, using Tr as prescribed by Lemmas 1 and 2.* This has at least two implications for research in meta-learning:

1. An important direction for meta-learning researchers to explore is the direct learning of $p^{\Omega}(f)$.
2. It may be that the problem of learning $p^{\Omega}(f)$ is just too complex, and that using existing learning algorithms as "gravitational" centers is an appropriate heuristic.

- There is a potentially interesting relationship between our framework and active learning. As per Lemma 2, given a learned $\hat{p}^{\Omega}(f)$, we should choose the MAP function or classification. However, "if I could see just one more training example...," the probabilities could be affected, helping to narrow down on the target function. This brings to mind active learning methods, like query by committee (Seung et al., 1992). The set of functions still consistent with the training data is the version space. Any OTS example for which we obtain the label will split the version space in half (in the binary case). But, splitting the version space in half may not be the best tack. One would like to choose the example with the best *expected* improvement, with respect to the target function.

- Meta-learning, as it is typically practiced, relies on similarity among learning domains (or training sets). The Ugly Duckling Theorem (Watanabe, 1985) states that even the seemingly simple concept of similarity requires (usually unstated) assumptions. It seems important also to clarify precisely the assumptions that meta-learning researchers make regarding meta-attributes and useful comparisons between learning domains.

7. Conclusion

Machine learning researchers make basic assumptions that circumvent the No Free Lunch theorem and render it mostly irrelevant to research in Machine Learning. There are various versions of these assumptions that vary in strength and justify different ways of conducting machine learning research. The most reasonable assumptions argue for meta-learning as the proper strategy for machine learning researchers to take, and certainly using learning to build learning programs is more consistent than using manual knowledge engineering to build learning programs. However, meta-learning researchers would be well advised to consider carefully whether and why the common strategies they pursue are best, and whether alternative strategies (such as direct learning of p^{Ω}) are better. Our goal in initiating this analysis was to provoke discussion at this workshop. Analyzing carefully the fundamental underpinnings of meta-learning certainly requires more and deeper thought.

References

Bensusan, H., & Giraud-Carrier, C. (2000). Discovering task neighbourhoods through landmark learning performances. *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), LNAI 1910* (pp. 325–330).

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification.* John Wiley & Sons, Inc. Second edition.

Goodman, N. (1946). A query on confirmation. *Journal of Philosophy*, *43*, 383–385.

Goodman, N. (1983). *Fact, fiction and forecast.* Cambridge, MA and London: Harvard University Press. 4 edition, First published in 1955.

Hume, D. (1740). *A treatise of human nature.* Edited by D.F. Norton and M.J. Norton, Oxford University Press, 2000.

Igel, C., & Toussaint, M. (2004). A no-free lunch theorem for non-uniform distributions of target functions. *Journal of Mathematical Modelling and Algorithms*, *3*, 313–322.

Kneale, W. (1949). *Probability and induction.* Clarendon Press, Oxford.

Michie, D., Spiegelhalter, D., & Taylor, C. (Eds.). (1994). *Machine learning, neural and statistical classification.* Ellis Horwood.

ML-List (1994). LCG discussion thread. *Machine Learning List*, *6*.

Pfahringer, B., Bensusan, H., & Giraud-Carrier, C. (2000). Meta-learning by landmarking various learning algorithms. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)* (pp. 743–750).

Rendell, L., & Cho, H. (1990). Empirical learning as a function of concept character. *Machine Learning*, *5*, 267–298.

Russell, S. (1986). Preliminary steps toward the automation of induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 477–484).

Schaffer, C. (1994). A conservation law for generalization performance. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 259–265).

Schumacher, C., Vose, M. D., & Whitley, L. (2001). The no free lunch and problem description length. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)* (pp. 565–570).

Seung, H., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Computational Learning Theory*, 287–294.

van Someren, M. (2001). Model class selection and construction: Beyond the procustean approach to machine learning applications. In G. Paliouras, V. Karkaletsis and C. Spyropoulos (Eds.), *Machine learning and its applications: Advanced lectures, lncs 2049*, 196–217. Springer-Verlag.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., & Soares, C. (2004). Using meta-learning to support data-mining. *International Journal of Computer Science Applications*, *I*, 31–45.

Watanabe, S. (1985). *Pattern recognition: Human and mechanical*. New York: Wiley.

Wolpert, D. (2001). The supervised learning no-free-lunch theorems. *Proceedings of the Sixth On-line World Conference on Soft Computing in Industrial Applications* (pp. 325–330).

Wolpert, D., & Macready, W. (1995). *No free lunch theorems for search* (Technical Report SFI-TR-95-02-010). Santa Fe Institute.

View publication stats