An Introduction to Metalearning

Christophe Giraud-Carrier
Introduction

Metalearning
  Theoretical Considerations
  Practical Considerations
  Rice’s Framework

The Practice of Metalearning
  Choosing the content of $A$
  Constructing the Training Metadata
  Choosing $f$
  Computational Cost of $f$ and $S$
  Choosing $p$
  Choosing the form of the output of $S$

Metalearning Systems
Machine Learning

- Machine learning focuses on accumulating experience about a specific learning task or application (e.g., medical diagnosis, fraud detection, etc.) so as to improve performance on it
- It is:
  - What we eat, drink and sleep
  - What makes the world go 'round
  - What we prescribe to anyone who would have it
- And YET...
The Shoemaker’s Children Syndrome

- Everyone is using Machine Learning!
  - Everyone, that is ...
  - Except us!
- Applied machine learning is guided mostly by hunches, anecdotal evidence, and individual experience
- If that is sub-optimal for our “customers,” is it not also sub-optimal for us?
- Shouldn’t we look to the data our applications generate to gain better insight into how to do machine learning?
- If we are not quack doctors, but truly believe in our medicine, then the answer should be a resounding YES!
We shall call *metadata* the type of data that may be viewed as being generated through the application of machine learning.

We shall call *metalearning* the use of machine learning techniques to build models from metadata.

Hence, metalearning is concerned with accumulating experience on the performance of multiple applications of a learning system.

Here, we will be particularly interested in the important problem of metalearning for algorithm selection.
Theoretical Considerations

- No Free Lunch (NFL) theorem / Law of Conservation for Generalization Performance (LCG)
  - When taken across all learning tasks, the generalization performance of any learner sums to 0
- Is Metalearning doomed?
NFL Revisited (1/4)

- Consider the space, \( \mathcal{F} \), of functions defined over \( \mathbb{B}^3 = \{0, 1\}^3 \)
- Assume that the instances of set \( Tr = \{000, 001, \ldots, 101\} \) are observed, and the instances of set \( Te = \mathbb{B}^3 - Tr = \{110, 111\} \) constitute the off-training set (OTS) test set

<table>
<thead>
<tr>
<th></th>
<th>Inputs</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
<th>( f_6 )</th>
<th>( f_7 )</th>
<th>( f_8 )</th>
<th>( f_9 )</th>
<th>( f_{10} )</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Set</strong></td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>0 0 1</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>0 1 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>0 1 1</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>1 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>1 0 1</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 0 0</td>
<td>0 0 0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td>1 1 0</td>
<td>0 0 1</td>
<td>1 1 0</td>
<td>0 1 1</td>
<td>0 1 1</td>
<td>0 0 0</td>
<td>0 ...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>1 1 1</td>
<td>0 1 0</td>
<td>1 0 1</td>
<td>0 1 0</td>
<td>1 0 1</td>
<td>0 1 0</td>
<td>1 ...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
NFL Revisited (2/4)

- NFL shows that, averaged over all \( f_1, f_2, \ldots, f_{256} \in \mathcal{F} \), the behavior on \( T_e \) of any learner trained on \( T_r \) is that of a random guesser.

- This result is rather intuitive:
  - Consider functions \( f_1 \) through \( f_4 \)
  - For all 4 functions, \( T_r \) is the same
  - Given any deterministic learner \( L \), the model induced by \( L \) from \( T_r \) is the same in all 4 cases
  - Since the associated \( T_e \)'s span all possible labelings of OTS, for any OTS instance any model will be correct for half the functions and incorrect for the other half
  - Argument is easily repeated across all such subsets of 4 functions, giving the overall result

Christophe Giraud-Carrier
An Introduction to Metalearning
NFL Revisited (3/4)

- NFL 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.

- All other things being equal, given that all we see is $Tr$ and its labeling, there is no rational reason to prefer one labeling of $Te$ over another
NFL Revisited (4/4)

- Crucial and most powerful contribution of NFL
  - Whenever a learning algorithm performs well on some function, as measured by OTS generalization, it must perform poorly on some other(s)
- Hence, building decision support systems for what learning algorithm works well where becomes a valuable endeavor
Fundamental Assumption

- We all hold deep-rooted, intuitive notions of bizarre functions
  - 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, i.e., 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
Practical Considerations

- When a designer introduces a novel classification algorithm, how does she position it in the existing algorithm landscape?
- When a practitioner is faced with a new task for which she seeks a high accuracy model, how does she know which algorithm to use?
Implications of NFL

Two views of the world

1. Closed Classification World Assumption (CCWA)
   - She assumes that all classification tasks likely to occur form some well-defined subset of the universe. As a designer, she shows that her novel algorithm performs better than others on that set. As a practitioner, she picks any of the algorithms that performs well on that set.

2. Open Classification World Assumption (OCWA)
   - She assumes no structure on the set of classification tasks. As a designer, she characterizes as precisely as possible the class of tasks on which her novel algorithm outperforms others. As a practitioner, she has some way of determining which algorithm(s) will perform well on her specific task(s).
ML Community Subconscious

- Widely-used approach consisting in benchmarking algorithm against well-known repositories (e.g., UCI) tends to implicitly favor the CCWA
- Yet, there is no known characterization of “real-life” classification tasks
Algorithm Design Issues

- To compound the problem for practitioners, most efforts in algorithm design seem to share the same oblivious pattern:
  1. They propose new algorithms that overcome known limitations. Yet, unless one accepts the CCWA, this simply shifts the original question of how to overcome the targeted limitations to the equally difficult question of determining what applications the proposed approach works well on.
  2. They “promote” new algorithms on the basis of limited empirical results, leaving the burden of proof to the users. It is not trivial to know how well any new approach will generalize beyond the problems it has been tested against so far.
Impact on Users

- Large number of learning algorithms, with comparatively little insight gained in their individual applicability
- Users are faced with a plethora of algorithms, and without some kind of assistance, algorithm selection can turn into a serious road-block for those who wish to access the technology more directly and cost-effectively
- End-users often lack not only the expertise necessary to select a suitable algorithm, but also the availability of many algorithms to proceed on a trial-and-error basis
- And even then, trying all possible options is impractical, and choosing the option that “appears” most promising is likely to yield a sub-optimal solution
DM Packages

- Commercial DM packages consist of collections of algorithms wrapped in a user-friendly graphical interface
  - Facilitate access to algorithms, but generally offer no real decision support to non-expert end-users
- Need an informed search process to reduce the amount of experimentation while avoiding the pitfalls of local optima
- Informed search requires metaknowledge
- Metalearning offers a robust mechanism to build metaknowledge about algorithm selection in classification
- In a very practical way, metalearning contributes to the successful use of Data Mining tools outside the research arena, in industry, commerce, and government
Rice’s Framework

- A problem $x$ in problem space $P$ is mapped via some feature extraction process to $f(x)$ in some feature space $F$, and the selection algorithm $S$ maps $f(x)$ to some algorithm $a$ in algorithm space $A$, so that some selected performance measure (e.g., accuracy), $p$, of $a$ on $x$ is optimal.
Framework Issues

- The following issues have to be addressed
  1. The choice of \( f \),
  2. The choice of \( S \), and
  3. The choice of \( p \).

- \( A \) is a set of base-level learning algorithms and \( S \) is itself also a learning algorithm

- Making \( S \) a learning algorithm, i.e., using metalearning, has further important practical implications about:
  1. The construction of the training metadata set, i.e., problems in \( P \) that feed into \( F \) through the characterization function \( f \),
  2. The content of \( A \),
  3. The computational cost of \( f \) and \( S \), and
  4. The form of the output of \( S \)
Choosing Base-level Learners

- No learner is universal
- Each learner has its own area of expertise, i.e., the set of learning tasks on which it performs well
- Select base learners with complementary areas of expertise
- Two issues in this choice:
  - Coverage
  - Size
Good Coverage

- Seek the smallest set of learners that is most likely to ensure a reasonable coverage
- Not trivial
  - Experiments on the space of binary classification tasks of 3 Boolean variables
  - From 26 applicable algorithms, a subset of 7 is sufficient to obtain maximal coverage
  - But 9 tasks still remain uncovered (i.e., none of the learners is better than chance on these)
- Recommendation:
  - Choose base learners have different biases by choosing representatives from varied model classes
  - The more varied the biases, the greater the coverage
Nature of Training Metadata

- **Challenge:**
  - Training data at metalevel = data about base-level learning problems or tasks
  - Number of accessible, documented, real-world classification tasks is small
Building Training Metadata

Two alternatives:
- Augmenting training set through systematic generation of synthetic base-level tasks
- View the algorithm selection task as inherently incremental and treat it as such

Recommendation:
- First approach is non-trivial and probably limited to well-defined data characteristics
- Second approach naturally adapts to reality, extending to new areas of the base level learning space only when tasks from these areas actually arise
Meta-examples

- Meta-examples are of the form \( < f(x), t(x) > \), where \( t(x) \) represents some target value for \( x \)
- By definition, \( t(x) \) is predicated upon \( p \), and the choice of the form of the output of \( S \)
- Focusing on the case of selection of 1 of \( n \):
  \[
  t(x) = \arg\max_{a \in A} p(a, x)
  \]
- Metalearning takes \( \{ < f(x), t((x)) > : x \in P' \subseteq P \} \) as a training set and induces a metamodel that, for each new problem, predicts the algorithm from \( A \) that will perform best
- Constructing meta-examples is computationally intensive
Choosing $f$

- As in any learning task, the characterization of the examples plays a crucial role in enabling learning
- Features must have some predictive power
- Four main classes of characterization:
  - Statistical and information-theoretic
  - Model-based
  - Landmarking
  - Learning Curves
Statistical and Information-theoretic Characterization

- Extract a number of statistical and information-theoretic measures from the labeled base-level training set
- Typical measures include number of features, number of classes, ratio of examples to features, degree of correlation between features and target, class-conditional entropy, skewness, kurtosis, and signal to noise ratio
- Assumption: learning algorithms are sensitive to the underlying structure of the data on which they operate, so that one may hope that it may be possible to map structures to algorithms
- Empirical results do seem to confirm this intuition
Model-based Characterization

- Exploit properties of a hypothesis induced on problem $x$ as an indirect form of characterization of $x$

- Advantages:
  1. Dataset is summarized into a data structure that can embed the complexity and performance of the induced hypothesis, and thus is not limited to the example distribution
  2. Resulting representation can serve as a basis to explain the reasons behind the performance of the learning algorithm

- To date, only decision trees have been considered, where $f(x)$ consists of either the tree itself, if the metalearning algorithm can manipulate it directly, or properties extracted from the tree, such as nodes per feature, maximum tree depth, shape, and tree imbalance
Landmarking

- Each learner has an area of expertise, i.e., a class of tasks on which it performs particularly well, under a reasonable measure of performance.

- Basic idea of the landmarking approach:
  - Performance of a learner on a task uncovers information about the nature of the task.
  - A task can be described by the collection of areas of expertise to which it belongs.

- A *landmark learner*, or simply a *landmarker*, a learning mechanism whose performance is used to describe a task.

- Landmarking is the use of these learners to locate the task in the expertise space, the space of all areas of expertise.
Computational Cost

- Necessary price to pay to be able to perform algorithm selection learning at the metalevel
- To be justifiable, the cost of computing $f(x)$ should be significantly lower than the cost of computing $t(x)$
- The larger the set $A$ and the more computationally intensive the algorithms in $A$, the more likely it is that the above condition holds
- In all implementations of the aforementioned characterization approaches, that condition has been satisfied
- Cost of induction vs. cost of prediction (batch vs. incremental)
Selecting on Accuracy

- Predictive accuracy has become the *de facto* criterion, or performance measure
- Bias largely justified by:
  - NFL theorem: good performance on a given set of problems cannot be taken as guarantee of good performance on applications outside of that set
  - Impossibility of forecasting: cannot know how accurate a hypothesis will be until that hypothesis has been induced by the selected learning model and tested on unseen data
  - Quantifiability: not subjective, induces a total order on the set of all hypotheses, and straightforward, through experimentation, to find which of a number of available models produces the most accurate hypothesis
Selecting on Other Criteria (1/2)

- There are a number of other aspects which have an impact on which model to select.
- In fact:
  - Empirical evidence suggests that for large classes of applications, most learners perform well in terms of accuracy.
  - Yet they often exhibit extreme variance along other dimensions.
- This (ab)use of accuracy is also a result of the assumption that predicting data values equates to predicting useful business outcomes.
  - Again, evidence suggests that this assumption may not always hold and other factors should be considered.
Selecting on Other Criteria (2/2)

- Other performance measures:
  - Expressiveness
  - Compactness
  - Computational complexity
  - Comprehensibility
  - Etc.

- These could be handled in isolation or in combination to build multi-criteria performance measures

- To the best of our knowledge, only computational complexity, as measured by training time, has been considered in tandem with predictive accuracy
Selection vs. Ranking

- Standard: single algorithm selected among $n$ algorithms
  - For every new problem, metamodel returns one learning algorithm that it predicts will perform best on that problem
- Alternative: ranking of $n$ algorithm
  - For every new problem, metamodel returns set $A_r \subseteq A$ of algorithms ranked by decreasing performance
Advantages of Ranking

- Ranking reduces brittleness
- Assume that the algorithm predicted best for some new classification problem results in what appears to be a poor performance
  - In the single-model prediction approach, the user has no further information as to what other model to try
  - In the ranking approach, the user may try the second best, third best, and so on, in an attempt to improve performance
- Empirical evidence suggests that the best algorithm is generally within the top three in the rankings
Although a valid intellectual challenge in its own right, metalearning finds its real raison d’être in the practical support it offers Data Mining practitioners.

Some promising implementations:

- MininMart
- Data Mining Advisor
- METALA
- Intelligent Discovery Assistant

Mostly prototypes, work in progress

- ExperimentDB