Assessing the Feasibility of Learning Biomedical Phenotypes via Large Scale Omics Profiles

Abstract

This paper applies the computational learning theory framework to elucidate the differences that distinguish hard bioinformatics learning tasks from easy. While most of the published predictive studies present the empirical error of a model used to learn a specific phenotype pattern given a group of subjects profiled by a recent omics measurement technology, very few explain why learning is feasible in some cases and infeasible in others. Our recent published results show that some tasks (such as predicting (sub)continental ancestral origins of individuals) are quite easy, while others (such as predicting the susceptibility to breast cancer) are extremely difficult. Our analysis suggests that the ancestral origin prediction problem is a case of realizable learning in the presence of many irrelevant features, which suggests that a training dataset with $\frac{1}{2} \left( \ln |H| + \ln \left( \frac{\delta}{8} \right) \right)$ samples would suffice for PAC learning this target concept. On the other hand, our analysis suggests that the breast cancer prediction problem appears a case of unrealizable learning from incomplete examples with relevant hidden features, and hidden subclasses, which suggests that at least a training dataset with $\max \left( d_1 \ln d_2 \times \ln \frac{1}{d_1 \varepsilon \ln (1 - 2\delta |\mathcal{H}|)} \right)$ samples is necessary for PAC learning this target concept in the worst case. The paper also discusses the effect of the number of irrelevant features, relevant hidden features, and hidden subclasses on the sample complexity of learning biomedical phenotypes – which is very relevant to our task involving high-throughput omics profiles. This paper can aid future omics researchers interested in predictive studies to estimate the necessary and sufficient number of training examples required for their predictive studies.

Introduction

In addition to the widespread interest of the biomedical community in conducting association and risk assessment studies on datasets generated by omics profiling, many biomedical researchers are now performing predictive studies. However, most of these predictive studies focus only on experimentally mix-and-matching different algorithms for pre-processing, feature selection, and learning, then conclude by presenting the empirical error of their model using an evaluation strategy such as cross validation or hold-out dataset. The omics field would benefit from analytical assessments that mathematically explain why learning is feasible in some cases but infeasible in others. This paper begins to fill this gap by introducing the computational learning theory framework to the biomedical community and using this framework to elucidate the differences between two recently published predictive studies, predicting breast cancer and ancestral origins, each using the omics profiles generated by genotyping the germline SNPs of samples on Affymetrix Human SNP 6.0 array [1-2]. These two studies had similar sample and feature sizes, but the prediction performance of the resulting predictive models were very different. As these studies lie at the extremes of the spectrum of predictability of biomedical phenotypes, analyzing them using the computational learning theory lens may help us to understand the limits of learnability of biomedical phenotypes. In the breast cancer prediction task [1], we utilized 696 samples (348 breast cancer cases and 348 apparently healthy controls) to learn a model that predicts whether a new subject will develop breast cancer, based on her SNP profile. Despite trying a wide range of biologically-aware and biologically-naïve (statistical) supervised learning approaches, we could never achieve an empirical error better than 0.4. In the ancestral origin prediction task [2], we utilized the international HapMap project Phase II and III datasets [3] to learn models that can predict an individual’s (sub)continental ancestral origins. While the breast cancer prediction had only marginal success, it was very easy to achieve empirical errors of less than 0.1 in the (sub)continental in predicting ancestral origins. For example, in the continental ancestral origin prediction problem, using 270 samples (1/3 in each continent), a single CART decision tree [5] with 3 internal nodes (SNPs), had an empirical error rate of 0.03, and an ensemble of 3 disjoint decision trees with 3-4 internal nodes (SNPs) each, achieved an empirical error rate of 0.
in form of disjoint small decision trees (ie, whose identified many equally good concepts/patterns, ancestral origins [2]. In any of these provides the information needed to identify one’s origins confirms that a small number of SNPs analysis on learning (sub)continental ancestral individual’s SNP profiling provides the means to It is well-known in the human genetics that an environmental, and lifestyle factors, we were Among these many different genetic, endogenous and exogenous sources), radiation factors include point mutations, SNPs, CNVs, lifestyle, and genetic factors. Heritable genetic of an interaction among many environmental, Breast Cancer Prediction Problem is a case of realizable learning in presence of viewpoint, the ancestral origin learning problem Therefore, it is not surprising that we could learn simple as a small decision tree is feasible. high-throughput SNP profiles using models as study suggests learning ancestral origins from features were disjoint); as these patterns were Table 1: Relevant Sample Complexity Bounds from the Computational Learning Theory Literature | Sample Complexity Upper-Bound | Sample Complexity Lower-Bound |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Realizable Learning</td>
<td>$\frac{1}{\varepsilon} (\ln</td>
</tr>
<tr>
<td>Unrealizable Learning</td>
<td>$\frac{1}{2\varepsilon} (\ln</td>
</tr>
<tr>
<td>Learning with a Fixed Labeling Noise Rate</td>
<td>$\frac{1}{2\varepsilon^2 (1-2\eta)^2} (\ln</td>
</tr>
</tbody>
</table>

Methods

**Ancestral Origin Prediction Problem**

It is well-known in the human genetics that an individual’s SNP profiling provides the means to identify his/her ancestral origins [6]. Our recent analysis on learning (sub)continental ancestral origins confirms that a small number of SNPs provides the information needed to identify one’s ancestral origins [2]. In any of these (sub)continental classification problems, we identified many equally good concepts/patterns, in form of disjoint small decision trees (ie, whose features were disjoint); as these patterns were accurate and diverse, we were able to increase the model accuracy by making an ensemble over these disjoint decision trees [7]. Our empirical study suggests learning ancestral origins from high-throughput SNP profiles using models as simple as a small decision tree is feasible. Therefore, it is not surprising that we could learn accurate predictor of ancestral origin from a small sample size in order of hundreds of samples. From the computational learning theory viewpoint, the ancestral origin learning problem is a case of realizable learning in presence of many irrelevant features.

**Breast Cancer Prediction Problem**

Like most cancers, breast cancer occurs because of an interaction among many environmental, lifestyle, and genetic factors. Heritable genetic factors include point mutations, SNPs, CNVs, and structural chromosome variations [8]. The major environmental and lifestyle risk factors include age, lack of childbearing or lack of breastfeeding, obesity, estrogen exposure (from endogenous and exogenous sources), radiation exposure, certain chemicals exposure, smoking, alcohol intake, and physical inactivity [9]. Among these many different genetic, environmental, and lifestyle factors, we were given only with SNPs in our breast cancer learning problem [1]. Furthermore, breast cancer is biologically heterogeneous disease, with a high degree of diversity between and within tumors as well as among cancer-bearing individuals and current molecular classifications -- based on clinical determinations of estrogen receptor status (ER), human epidermal growth factor receptor 2 status (HER2), or proliferation rate status (PR) -- suggest a minimum of four distinct biological subtypes for breast cancer [10]. However, these distinctions are ignored in our dataset and all these subclasses are merged into the single “breast cancer” label. From the computational learning theory viewpoint, the breast cancer learning problem appears a case of unrealizable learning with many irrelevant features, relevant hidden features, and hidden subclasses.

**Computational Learning Theory**

Computational learning theory is a subfield of machine learning whose theorems explain the required computational and sample complexity of learning a pattern [11]. We consider both upper and lower sample complexity bounds within the PAC learning setting [12-14]. Sample complexity upper-bound for PAC learning a concept class C from the hypothesis class H, $m_U(C, H, \varepsilon, \delta)$, is the number of training examples that is sufficient for finding an hypothesis $h \in H$ that is $(\varepsilon, \delta)$-close to the target concept $c \in C$ -- ie, for any $c \in C$, we can identify an $h \in H$ whose error is within $\varepsilon$ of $c$’s error, with probability at least $1-\delta$. Sample complexity lower-bound for PAC learning C from H, $m_L(C, H, \varepsilon, \delta)$, is the number of training examples that are necessary for finding an $h \in H$ that is $(\varepsilon, \delta)$-close to $c \in C$. Learning is feasible given at least $m_U(C, H, \varepsilon, \delta)$ training instances, and is infeasible given less training instances than $m_L(C, H, \varepsilon, \delta)$, in the worst case -- ie, for the
worst possible choice of the target concept and the worst possible distribution of training examples.

Table 1 presents some upper-bound and lower-bound for PAC learning various learning problems pertinent for analyzing the (in)feasibility of our learning tasks, including results related to realizable learning, unrealizable learning, and learning with a fixed labeling noise rate:

A hypothesis class \( H \) over the input space \( X \) for learning the target concept \( c \) is realizable if the optimal Bayes error of \( H \) equals zero and is unrealizable if it is greater than zero. The optimal Bayes error of learning the target concept \( c \) using the hypothesis class \( H \) equals \( L_H = \inf_{h \in H} \{ \text{error}_D(h) = \Pr_{X \sim D}[c(x) \neq h(x)] \} \).

In learning under a fixed labeling noise rate, the label of each instance is flipped (ie, reversed) randomly with probability \( \eta < 0.5 \) [19]. This problem is proved to be a type of probabilistic concept learning. Probabilistic concept learning is a class of supervised learning problems in which the concept to be learned may exhibit uncertain or probabilistic behavior. Thus, the same instance may sometimes be classified as a positive example and sometimes as a negative example. An example of probabilistic concept learning is predicting tomorrow's weather as accurately as possible via measuring a small number of presumably relevant features, such as the current temperature, barometric pressure, and wind speed and direction. A possible forecast would be of the form "chances for rain tomorrow are 70%." The next day it either rains or it does not rain [20].

In learning from incomplete examples or learning with relevant hidden features, there is an underlying deterministic target concept, but some of the relevant variables are invisible to the learning algorithm, resulting in apparent probabilistic behavior [22]. This too is a type of probabilistic concept learning [20]. That is true for the breast cancer task, as the learner has access only to SNPs, but not other heritable factors, nor any lifestyle nor environmental factors. If we consider an extended training dataset that includes enumeration of the different values of the hidden features, the problem of learning with relevant hidden features would be translated into the problem of learning in the presence of fixed labeling noise. Thus, the sample complexity bounds of this class of learning problems considering an extended hypothesis class over visible and hidden features, would be the same as the case of learning under a fixed labeling noise as presented in Table 1.

**Results**

**Ancestral Origin Prediction Problem**

Learning ancestral origins pattern from SNP profiles is easy as the sample complexity upper-bound for PAC learning this concept suggests. Here, we only explain the case of continental population identification problem. However, the same sort of analysis can explain the story of the subcontinental population identification problems. The sample complexity upper-bound for PAC learning a target concept from the hypothesis class \( H \) in the realizable learning case is \( \frac{1}{\varepsilon}(\ln |H| + \ln \frac{1}{\delta}) \). The size of the hypothesis class of 3-node decision trees from \( p = 611146 \) features (SNPs), when there are 3 labels (African, European, and Asian) is \( \leq C_3 \times 2^3 \times 3^3 \times \left( \binom{611146}{3} \right) \leq 9720 \times 611146^3 \). Here \( C_3 = 5 \) is the number of binary decision trees with 3 nodes which equals the 3rd Catalan number. Therefore, considering \( \varepsilon = 0.05 \) and \( \delta = 0.01 \), the sample complexity upper-bound for learning this problem would be \( \frac{1}{0.05}(\ln(9720) + 3 \times \ln(611146) + \ln(\frac{1}{0.01})) \approx 1075 \). This implies that even in the worst case choices of the target concept and distribution of the training instances, having 1075 instances is sufficient for PAC learning this hypothesis class. However, we found that 270 instances suffice for PAC learning the continental ancestral origin pattern.

**Breast Cancer Prediction Problem**

Learning breast cancer from SNP profiles is tricky as the sample complexity lower-bound for PAC learning this target concept suggests. The sample complexity lower-bound for PAC learning the breast cancer target concept combining unrealizable learning and learning with relevant hidden features bounds using the finite hypothesis class \( H \) over visible features and the finite hypothesis class \( H' \) over visible and hidden features in the unrealizable learning case would be \( \max\left( \frac{\ln \left( \frac{1}{4\varepsilon^2} \cdot \frac{d_1 - 1}{8} \cdot \frac{\ln \left( \frac{1}{4\varepsilon^2} \cdot \frac{d_2}{48^5 \cdot e(1 - 24\varepsilon^2)} \right)}{2} \right)}{d_2} \right) \).

Based on the below-mentioned analysis we demonstrate that even when we consider a very small hypothesis class such as conjunctions of \( r \) out of \( p \) features, sample complexity lower-bound is very large. We find the sample
complexity lower-bound of \( \max \left( \frac{0.3}{4 \times 0.05^2} \times \ln \left( \frac{1}{4 \times 0.05} \right) \right) \approx 190 \). The sample complexity lower-bound increases if the number of hidden variables increases, the optimal Bayes error increases, and the sample complexity bounds increase consequentially with a squared rate. Therefore, we can judge that hidden variables could have a dramatic effect on the sample complexity bounds.

In many real-world learning problems, such as our breast cancer learning problem from SNP profiles, there are hidden subclasses in the labels provided for the learner as the disease. Existence of these implicit subclasses in fact increases the complexity of the target concept \( c \). This motivates using a more complex hypothesis class such as \( m \)-term r-DNF formulas out of \( p \) variables (each of \( m \) term represent a subclass) instead of conjunctions over \( r \) out of \( p \) features used in the Results section. Considering this specific example, at least in cases which we choose the hypothesis class to be the same as concept class, we observe that both the sample complexity upper-bound and lower-bound increase linearly with the increase of the number of hidden subclasses, considering the target concept to be a \( m \)-term r-DNF formula. Denote that \( m_d \propto \ln |H| \propto m \times r \times \ln(2p) \) and \( m_r \propto d \propto m \times r \times \log_2(p) \) [23].

**References**
