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 $\frac{1}{\varepsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$ 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 $\max\left(\frac{L_H}{4\varepsilon^2} \times \frac{d_1 - 1}{8}, \frac{L_H}{4\varepsilon^2} \times \ln\frac{1}{4\delta}, \frac{d_2}{\varepsilon(1-2L_{H'})^2}\right)$ 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.

Table 1: Relevant Sample Complexity Bounds from the Computational Learning Theory Literature – H: hypothesis class; d: VC dimension of H; L_H: optimal Bayes error rate of H; ε: estimation parameter; δ: confidence parameter; η: fixed labeling noise rate.

	Sample Complexity Upper-Bound	**Sample Complexity Lower-Bound**
Realizable Learning	$\frac{1}{\varepsilon}\left(\ln\lvert H\rvert + \ln\frac{1}{\delta}\right) = O\left(\frac{1}{\varepsilon}\left(\ln\lvert H\rvert + \ln\frac{1}{\delta}\right)\right)$ [15]	$\frac{1}{\varepsilon}\left(\max\left(\frac{d-1}{4}, \frac{\ln\frac{1}{\delta}}{8}\right)\right) = \Omega\left(\frac{1}{\varepsilon}\left(d + \ln\frac{1}{\delta}\right)\right)$ [16]
Unrealizable Learning	$\frac{1}{2\varepsilon^2}\left(\ln\lvert H\rvert + \ln\frac{2}{\delta}\right) = O\left(\frac{1}{\varepsilon^2}\left(\ln\lvert H\rvert + \ln\frac{1}{\delta}\right)\right)$ [17]	$\frac{L_H}{4\varepsilon^2}\left(\max\left(\frac{d-1}{8}, \ln\frac{1}{4\delta}\right)\right) = \Omega\left(\frac{L_H}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right)\right)$ [18]
Learning with a Fixed Labeling Noise Rate	$\frac{1}{2\varepsilon^2(1-2\eta)^2}\left(\ln\lvert H\rvert + \ln\frac{2}{\delta}\right) =$ $O\left(\frac{1}{\varepsilon^2(1-2\eta)^2}\left(\ln\lvert H\rvert + \ln\frac{1}{\delta}\right)\right)$ [19]	$\frac{d}{\varepsilon(1-2\eta)^2} = \Omega\left(\frac{d}{\varepsilon(1-2\eta)^2}\right)$ [21]

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, $m_U(C, H, \epsilon, \delta)$, is the number of training examples that is *sufficient* for finding an hypothesis h∈H that is (ε, δ)-close to the target concept c∈C -- ie, for any c∈C, we can identify an h∈H whose error is within ε of c's error, with probability at least 1-δ. Sample complexity lower-bound for PAC learning C from H, $m_L(C, H, \epsilon, \delta)$, is the number of training examples that are *necessary* for finding an h∈H that is (ε, δ)-close to c∈C. Learning is feasible given at least $m_U(C, H, \epsilon, \delta)$ training instances, and is infeasible given less training instances than $m_L(C, H, \epsilon, \delta)$, 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 $L_H = \inf_{h \epsilon H} \{error_{c,D}(h) = Pr_{X \epsilon D}[c(x) \neq h(x)]\}$.

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 $\frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$. 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 $\leq C_3 \times 2^3 \times 3^4 \times \binom{611146}{3} \leq 9720 \times 611146^3$ (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 $\frac{1}{0.05}\left(\ln(9720) + 3 \times \ln(611146) + \ln(\frac{1}{0.01})\right) \approx 1075$. 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 $\max\left(\frac{L_H}{4\varepsilon^2} \times \frac{d_1 - 1}{8}, \frac{L_H}{4\varepsilon^2} \times \ln\frac{1}{4\delta}, \frac{d_2}{\varepsilon(1-2L_{H'})^2}\right)$. 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 $\max\left(\frac{0.3}{4\times0.05^2}\times\frac{190-1}{8},\frac{0.3}{4\times0.05^2}\times\ln\left(\frac{1}{4\times0.01}\right),\frac{380}{0.05\times(1-2\times0.25)^2}\right) \approx \max(713,96,30400) = 30400$. Learning would be infeasible having a training dataset with less training examples than the sample complexity lower-bound in the worst case.

- p = 506836 (the number of unfiltered input features in the breast cancer learning problem).
- r_1 = 10 (the number of terms in a conjunction over visible features).
- $d_1 = r_1 \times \log_2 p$ = 190 (VC dimension of the hypothesis class of conjunctions of r_1 out of p features) [23].
- L_H = 0.3 (the optimal Bayes error of the hypothesis class of conjunctions of r_1 out of p features).
- h = 10 (the number of relevant hidden features).
- $r_2 = r_1 + h$ = 20 (the number of terms in a conjunction over visible and hidden features).
- $d_2 = r_2 \times \log_2(p + h)$ = 380 (VC dimension of the hypothesis class of conjunctions of r_2 out of p+h features) [23].
- $L_{H'}$ = 0.25 (the optimal Bayes error of the hypothesis class of conjunctions of r_2 out of p features).
- ε = 0.05 (estimation parameter).
- δ = 0.01 (confidence parameter).

Conclusion

Learning disease-associated phenotypes from omics profiles usually involves dealing with one or more of the following three challenges: 1) many irrelevant features exist in the problem input domain, 2) the learner does not have access to some very relevant features, and 3) the learner does not have access to the hidden subclasses in the class labels of the instances. It would be beneficial if we could estimate and then compare how the sample complexity upper-bounds and lower-bounds of learning problems vary given these characteristics.

Learning with many irrelevant features means the training dataset offered to the learner encompasses many features that are irrelevant to the target concept. The sample complexity upper-bound and lower-bound for learning with r relevant and p-r irrelevant features, given the Boolean functions hypothesis class, are $O\left(\frac{1}{\varepsilon}\left(2^r\times\ln 2+r\times\ln p+\ln\frac{1}{\delta}\right)\right)$ and $\Omega\left(\frac{1}{\varepsilon}\left(2^r+r\times\ln 2\times\ln p+\ln\frac{1}{\delta}\right)\right)$ [24]. As these bounds suggest, the presence of many irrelevant features does not make the learning task substantially more difficult, at least in terms of the number of examples needed for learning, since these sample complexity bounds grows only logarithmically in the number of irrelevant features. However, depending on the algorithm used, the computational complexity might be an issue while dealing with many irrelevant features.

The sample complexity bounds of learning a target concept in presence of relevant hidden features is dependent on the optimal Bayes error of learning the target concept via the hypothesis class over visible and hidden features by a factor of $\frac{1}{(1-2L_{H'})^2}$. If the number of hidden variables increases, the optimal Bayes error increases, and the sample complexity bounds increase consequently 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 $m_U \propto \ln|H| \propto m\times r\times\ln(2p)$ and $m_L \propto d \propto m\times r\times\log_2(p)$ [23].

References

1.Hajiloo M, et al.: **Using genome wide single nucleotide polymorphism data to learn a model for breast cancer prediction**, BMC Bioinformatics 2013, **14**(S13): S3.

2.Hajiloo M, et al.: **ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction**, BMC Bioinformatics 2013, **14**(1): 61.

3.The International HapMap Consortium: **The International HapMap project**. Nature 2003, **426**: 89-796.
4.Dietterich TG: **Ensemble methods in machine learning**. Lecture Notes in Computer Science 2000, **1857**:1-15.
5.Quinlan JR: **Induction of decision trees**. Machine Learning 1986, **1**:81-106.
6. Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS: **Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms**. BMC genomics 2007, **8**(1): 68.
7.Kuncheva LI, Whitaker CJ: **Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy.** Machine Learning 2003, **51**(2)**:**181-207.
8.Cho WC: An Omics Perspective on Cancer Research. New York, NY: Springer 2009.
9.Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. Nature 2009, **461:**747-753.
10. Bertucci F, Birnbaum D: **Reasons for breast cancer heterogeneity**. J Biology 2008, **7**(2):6.
11. Angluin D: **Computational learning theory: survey and selected bibliography**. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* 1992, pp. 351-369.
12. Valiant LG: **A theory of the learnable.** *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of* computing 1984.
13. Vapnik V, Chervonenkis A: **On the uniform convergence of relative frequencies of events to their probabilities**. *Theory of Probability and its Applications* 1971, **16**(2): 264-280.
14. Kearns M, Vazirani UV: *An introduction to computational learning theory*. The MIT Press 1994.
15. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK: **Occam's razor**. Information processing letters 1987, **24**(6): 377-380.
16. Ehrenfeucht A, Haussler D, Kearns M, Valiant L: **A general lower bound on the number of examples needed for learning**. Information and Computation 1989, **82**(3): 247-261.
17. Haussler D: **Decision theoretic generalizations of the PAC model for neural net and other learning applications**. Information and computation 1992, **100**(1): 78-150.
18. Devroye L, Lugosi G: **Lower bounds in pattern recognition and learning**. Pattern recognition 1995, **28**(7): 1011-1018.
19. Angluin D, Laird P: **Learning from noisy examples**. Machine Learning 1988, **2**(4): 343-370.
20. Kearns MJ, Schapire RE: **Efficient distribution-free learning of probabilistic concepts**. J Computer and System Sciences 1994, **48**(3):
21. Simon HU: **General bounds on the number of examples needed for learning probabilistic concepts**. J Computer and System Sciences 1996, **52**(2): 239-254.
22. Schuurmans D, Greiner R: **Learning to classify incomplete examples**. Computational Learning Theory and Natural Learning Systems, 1997, **4**: 87-105.
23. Littlestone N: **Learning quickly when irrelevant attributes abound**. Machine Learning 1988, **2**(4): 285-318.
24. Almuallim H, Dietterich TG: **Learning Boolean concepts in the presence of many irrelevant features**. Artificial Intelligence 1994, **69**(1): 279-305.