A Theoretical Framework for Learning from a Pool of Disparate Data Sources

Shai Ben-David
Computer Science
Technion, Haifa 32000, Israel
and Cornell University
Ithaca, NY 14853
shai@cs.cornell.edu

Johannes Gehrke
Computer Science
Cornell University
Ithaca, NY 14853
johannes@cs.cornell.edu

Reba Schuller
Department of Mathematics
Cornell University
Ithaca, NY 14853
ras51@cornell.edu

ABSTRACT

Many enterprises incorporate information gathered from a variety of data sources into an integrated input for some learning task. For example, aiming towards the design of an automated diagnostic tool for some disease, one may wish to integrate data gathered in many different hospitals. A major obstacle to such endeavors is that different data sources may vary considerably in the way they choose to represent related data. In practice, the problem is usually solved by a manual construction of semantic mappings and translations between the different sources. Recently there have been attempts to introduce automated algorithms based on machine learning tools for the construction of such translations.

In this work we propose a theoretical framework for making classification predictions from a collection of different data sources, *without creating explicit translations between them*. Our framework allows a precise mathematical analysis of the complexity of such tasks, and it provides a tool for the development and comparison of different learning algorithms. Our main objective, at this stage, is to demonstrate the usefulness of computational learning theory to this practically important area and to stimulate further theoretical and experimental research of questions related to this framework.

1. INTRODUCTION

While the world has accumulated an astronomical amount of data, use of this data is significantly inhibited by our inability to efficiently integrate data from multiple sources. The strength of the XML standardization movement[1] is a good indicator of the significance of this problem. In addition to motivating standardization for future data collection, this problem has inspired investigation into methods for integrating non-standardized data. Much of this work has focused on data integration systems, which allow access to multiple data sources through a single mediated schema. However, the problem of creating semantic mappings (i.e., mappings which preserve the meaning of the data) between the original schemas and the mediated schema remains a serious bottleneck.

We address the problem of data integration for the purpose of classification prediction. In this setting, the data has the form of a set of attributes and a classification ('yes' or 'no'), and one desires the ability to predict the classifications of new attribute combinations based on several small sets of training examples. We show that there are quite natural scenarios in which this can be done without the creation of explicit semantic mappings. That is, the information gathered from multiple data sources (without knowledge of the semantic mappings between them) can be utilized to guarantee better predictions than any algorithm could have obtained by considering any one of these sources on its own (with explicit knowledge of its semantics).

As an example, consider a hospital hoping to improve its methods for determining which patients diagnosed with pneumonia are at high enough risk to warrant hospitalization. While standard machine learning tools allow them to use relevant data (patient histories for patients previously diagnosed with pneumonia) collected by the hospital to make these predictions, it is known that larger sets of data guarantee a greater likelihood of more accurate prediction. Similar data sets from many other hospitals are available, and it would be ideal to augment the data set with data from them; however, use of these data sets is hindered by the use of different database schemas by different hospitals. Previously, one would have to first determine the different schemas and translate all the data into one unified database; however, we offer a new alternative: in the sequel, we describe a technique for using the raw databases (without modification) to improve prediction.

1.1 Our Model

We model a situation in which some fixed probability distribution generates database entries (examples) and their labels (classifications). A learner has access to a collection of databases with labeled entries, sampled i.i.d. from this distribution. However, each database presents these entries using its own representation schema. The task of the learner is to predict, based on these labeled examples, the labels of

[1]The goal of the XML standardization movement is for every industry to have a set of standard data structures for exchange of information.

new entries in one of the databases. While the identities of the data sources and the set of potential representation schemas used by the database are known to the learner, the learner does not know which schema corresponds to which database.

1.2 Results

We demonstrate the utility of computational learning theory to the problem of learning from disparate data sources. We present a meta-algorithm for incorporating information from data sources that are encoded in ways unknown to the learner, in a way that achieves classification prediction that is provably better than what can be achieved by algorithms that base their predications on any one of the data sources alone. Our main results show that, having access to sufficiently many disparate sources, classification prediction becomes significantly more effective than learning from any one of the sources alone, *without determining the explicit semantic mappings between them*, and in particular, without the need to apply semantic knowledge about the different schemas used by these data sets. In this work, we focus on the information complexity of the prediction task, setting aside the computational complexity aspects of this problem.

More specifically, fix a set of potential hypothesis (classification) functions $\mathbb{H}$, and assume that the semantic translation functions, used to encode the data in each of the data sets, come from some family of functions $\mathcal{F}$. For example, a typical such family $\mathcal{F}$ might include permutations (to allow for different orderings of attributes), as well as linear functions acting on individual attributes (to allow for disparity in the units in which attributes are measured.) Let $D_1, \ldots, D_n$ be a sequence of sample sets, each consisting of entries from one database and the corresponding classifications of those entries. We propose the following meta-algorithm:

> On input of $D_1, \ldots, D_n$, search for some $h \in \mathbb{H}$ and functions $f_1, \ldots, f_n \in \mathcal{F}$, that minimize the average of the empirical errors of $h \circ f_i$ on D_i. To predict the label on a new entry, represented using the scheme of some data set D_i, apply the resulting $h \circ f_i$ to this new entry.

We provide formal tools to analyze the prediction success of such an algorithm. Our results provide guaranteed error bounds in terms of combinatorial parameters that depend on the richness of the hypothesis class and of the family of data transformations, and on the relationship between these two classes of functions.

Since we do not restrict the label-generating distribution in any way, the results we get are relative to the best predictor within our class of hypothesis. This is in line with the common agnostic learning (or statistical regression) framework.

The rest of the paper is organized as follows. We begin, in Section 2, by discussing our learning approach for a concrete example, namely the task of learning Euclidean rectangles for data sets that are distorted by unknown linear shifts. We prove that in this case, our algorithmic approach allows the use of the disparate data sources almost as if they were all undistorted and merged into a large unified data set. Preparing the ground for a more general analysis, we provide in Section 3 a brief overview of some fundamental concepts from computational learning theory which will be used throughout the paper. We then go on to outline , in 3.3, a relevant result from [1] bounding the generalization error of learning in the framework of multiple related tasks. In section 4 we prove error bounds for a general case. These bounds are formulated in terms of combinatorial parameters that measure the information complexity of our learning task. Section 5 discusses related work and, finally, we offer concluding remarks in section 6.

2. LEARNING RECTANGLES FROM SHIFTED DATA SETS

Let us begin by considering a concrete example. Let the database entries consist of points in the Euclidean space $\mathbb{R}^d$ (for some fixed d), and assume that the the classification of a point is determined by some axis aligned rectangle (unknown to the learner). That is, a point is labeled 1 if it lies within this rectangle, and 0 otherwise.

Assume that the different database schemas are obtained by shifts (of the points) of $\mathbb{R}^d$. I.e., for each database i there is some vector v^i, and a point $x = (x_1, \ldots, x_d)$ is represented in the i'th database as $x + v^i = (x_1 + v^i_1, \ldots, x_d + v^i_d)$.

We will show that the use of such extra databases helps in learning the side lengths of the rectangle used by the labeling algorithm (although the algorithm is not provided with the shifting vectors v^i)..

Informally, the learning algorithm is as follows: For each database, D_i, find smallest rectangle, R_i, containing all points in that database labeled 1. Output $(r_1, \ldots, r_d)$, where r_j is the maximal length of the j'th side over all the R_i's.

We show that with this algorithm, the learner's predictions based on n disparate data sets of size m are as good as could be guaranteed based on a single data set of size $n(m-c)$, where c is a constant depending on the Euclidean dimension and the desired prediction accuracy.

We now state formally the algorithm described above and analyze its performance.

Algorithm: For each data set D_i, and each coordinate j, $1 \leq j \leq d$, take

$$m_{i,j} = \max\{|x_j - y_j| : ((x_1, \ldots x_d), 1), ((y_1, \ldots y_d), 1) \in D_i\},$$

and take

$$r_j = \max_{1 \leq i \leq n} (m_{i,j}).$$

Theorem 2.1. *Let P be any probability distribution over $\mathbb{R}^d$, let R be an axis aligned rectangle in $\mathbb{R}^d$, and let $V^1, \ldots, v^n$ be d dimensional vectors. Let $D_1, \ldots, D_n$, be sets of labeled points, so that D_i is generated by sampling i.i.d. via P points $x \in \mathbb{R}^d$, and putting in D_i the labeled point $(x + v^i, R(x))$ (where $R(x) = 1$ in $x \in R$ and 0 otherwise).*

For any $\epsilon > 0$, if the algorithm, $\mathcal{A}$, described above is applied to $D_1, \ldots, D_n$, and if the size of each data set D_i exceeds m,

> *with probability exceeding $d\left(2\left(1 - \frac{\epsilon}{2d}\right)^m\right)^n$, $\mathcal{A}$ outputs a vector of side lengths $\bar{r}$ that has error less than ϵ.*

(Where the error of a rectangle r relative to P and R is the probability that a P-random point will be classified differently by r and R).

Note that this generalization guarantee is as good as the one that is obtained by the standard considerations [8] from a single training data set of size $n(m-c)$, where c is a

constant dependent on ϵ and the Euclidean dimension d. That is, for the purpose of learning the side lengths of a target rectangle, the distortion by unknown shifts can be overcome by the learning algorithm; beyond a certain "orientation constant" per database, each example contains as much information as if it had not been shifted at all!

PROOF. *Definition:* Given rectangle $\mathbf{r} = [v_1, v_1 + s_1] \times \ldots \times [v_d, v_d + s_d]$, distribution P on $\mathbb{R}^d$, and $0 \leq \epsilon \leq 1$, we define $\mathbf{r}_j(\epsilon) =$

$$[v_1, v_1 + s_1] \times \ldots \times [v_j, v_j + \delta] \times \ldots \times [v_d, v_d + s_d],$$

where δ is such that $P(x \in \mathbf{r}_j(\epsilon)) = \epsilon$, and $\mathbf{r}'_j(\epsilon) =$

$$[v_1, v_1 + s_1] \times \ldots \times [v_j + s_j - \delta', v_j + s_j] \times \ldots \times [v_d, v_d + s_d],$$

where δ' is such that $P\left(x \in \mathbf{r}'_j(\epsilon)\right) = \epsilon$.

Informally, $\mathbf{r}_j(\epsilon)$ and $\mathbf{r}'_j(\epsilon)$ are simply slices of thickness δ, δ' (in coordinate j) off of opposite sides of $\mathbf{r}$ such that for a point (x, b) chosen according to P, x has probability ϵ of lying within each of the resulting slices.

Consider the target rectangle, $\mathbf{r}$. If for some j, $1 \leq j \leq d$, some D_i contains (the appropriate shift of) a point from both $\mathbf{r}_j(\frac{\alpha}{2})$ and $\mathbf{r}'_j(\frac{\alpha}{2})$, then the error contributed by coordinate j is less than α. Thus, the total probability of getting more than error α from an inaccuracy in coordinate j is at most $\left(2\left(1 - \frac{\alpha}{2}\right)^m\right)^n$. In order for the total error to exceed ϵ, there must exist $\alpha_1, \ldots \alpha_d$, with $\sum_{i=1}^d \alpha_j > \epsilon$, such that the error contributed by coordinate j is at least α_j. In this case, some α_j is such that $\alpha_j > \epsilon/d$, so there must be some coordinate which contributes an error of at least ϵ/d. The probability of this occurring is at most $d\left(2\left(1 - \frac{\epsilon}{2d}\right)^m\right)^n$. □

Our main objective in this work is to investigate the extent to which such performance can be obtained in more general situations. In particular when the labeling is nondeterministic, and for arbitrary prediction functions and other families of data transformations (in place of the rectangles and shifts).

3. BACKGROUND

In this section we introduce some standard concepts and tools of Computational Learning Theory. For all this and more, see [2].

The standard approach to classification prediction problems is to begin with some set $\mathbb{H}$ (called a hypothesis space) of possible classification functions, i.e. subsets of the domain $\mathcal{X}$ (or equivalently, their characteristic functions, for $h \in \mathbb{H}$, $\chi_h : \mathcal{X} \to \{0, 1\}$

$$\chi_h(x) = \left\{ \begin{array}{l} 0 \text{ if } x \notin h \\ 1 \text{ if } x \in h \end{array} \right\}.)$$

Then one goes about selecting a good hypothesis from $\mathbb{H}$ based on available data.

3.1 VC-Dim and Information Complexity

For hypothesis space $\mathbb{H}$ over domain $\mathcal{X}$, we have the following definitions:

Definition 3.1. *For any $A \subseteq \mathcal{X}$, define*

$$\Pi_{\mathbb{H}}(A) = |\{h \cap A \ : \ h \in \mathbb{H}\}|.$$

Definition 3.2. *For natural number m, define*

$$\Pi_{\mathbb{H}}(m) = \max\{\Pi_{\mathbb{H}}(A) \ : \ |A| = m\}.$$

In words, $\Pi_{\mathbb{H}}(m)$ is the maximal number of subsets of S that can be obtained by $\mathbb{H}$ on any set S of size m.

Definition 3.3.

$$VC\text{-}dim(H) = \max\{m : \Pi_{\mathbb{H}}(m) = 2^m\}.$$

The VC-dimension is a measure of the complexity of a hypothesis space. The following, known as Sauer's lemma, demonstrates how the VC-dimension of a hypothesis space controls the number of dichotomies induced by H on a set of *any* size, m.

Theorem 3.4. *For all $m \in \mathbb{N}$, $\Pi_H(m) \leq (em/d)^d$, where $d = VC\text{-}dim(H)$.*

Furthermore, the VC-dimension gives a bound on the information complexity of ordinary classification prediction, i.e. how large a sample is needed to ensure that a hypothesis that performs well on the sample data is likely to perform nearly as well on new data.

We measure the empirical error of a hypothesis h on a sample (or training) set T by

$$\hat{Er}^T(h) = \frac{|\{(x, b) \in T \ : \ h(x) \neq b\}|}{|T|}.$$

The true error of h on distribution P is

$$Er^P(h) = P\{(x, b) \in X \times \{0, 1\} \ : \ h(x) \neq b\}.$$

Theorem 3.5. *If T is a random sample from distribution P on $\mathcal{X} \times \{0, 1\}$ with*

$$|T| \geq (64/\epsilon^2)[\log(4/\delta) + 2\, VC\text{-}dim(\mathbb{H}) \log(12/\epsilon)]$$

then for any $h \in \mathbb{H}$, with probability at least $1 - \delta$,

$$\left| Er^P(h) - \hat{Er}^T(h) \right| \leq \epsilon.$$

3.2 Standard Learning Frameworks

Before one can state formal generalization results, a distinction is usually made between two cases; The realizable case, often called *the PAC framework*, and the general (non-realizable) case, *the Agnostic framework*. In the PAC framework, it is assumed that that some hypothesis $h \in \mathbb{H}$ is a perfect classifier, i.e., $Er^P(h) = 0$. (Such an h is called a target function.) In this framework, it is reasonable to ask that a learner find a hypothesis h such that $Er^P(h)$ is small. While this assumption allows for nice clean analysis of learning problems, it is rarely true in real-world problems. In contrast, the agnostic framework allows the distribution to be arbitrary. In this framework, it is standard to judge the success of the learner relative to the best hypothesis, h_o in the given hypothesis space; thus, if a learning algorithm produces hypothesis h, we are satisfied with a result regarding the probability that $|Er^P(h) - Er^P(h_o)| < \epsilon$.

With the exception of section 2, we work in the agnostic framework.

3.3 A Related Problem

Our general results rely on work of Baxter [1] on multitask learning, which we introduce in this section.

Section 3.4 of [1] considers the following problem: Suppose we have n probability distributions, $P_1, \ldots, P_n$, on $\mathcal{X} \times \{0,1\}$, and for each i, we have sample set D_i generated by sampling m times from $X \times \{0,1\}$ according to P_i. Given a boolean hypothesis space family $\mathbf{H}$ over $\mathcal{X}$, if we choose hypothesis h_i to approximate D_i, such that $h_1, \ldots, h_n$ are all from some single $H \in \mathbf{H}$ how well does this sequence of hypotheses generalize to $P_1, \ldots, P_n$?

In order to address this problem, [1] introduces the analog to the VC-dimension that appears as our definition 3.7.

Notation. For function $g : Y \to Z$ and $\overline{y} = (y_1, \ldots, y_n) \in Y^n$, $\overline{g}(\overline{y})$ will denote $(g(y_1), \ldots, g(y_n)) \in Z^n$.

Definition 3.6. $\Pi_{\mathbf{H}}(n,m) =$

$$\max_{\overline{x}_1,\ldots,\overline{x}_n \in X^m} \left| \left\{ \begin{bmatrix} \overline{h_1}(\overline{x_1}) \\ \vdots \\ \overline{h_l}(\overline{x_n}) \end{bmatrix} : \exists H \in \mathbf{H} \text{ with } h_1, \ldots, h_n \in H \right\} \right|$$

Definition 3.7. $d_{\mathbf{H}}(n) = \max\{m \ : \ \Pi_{\mathbf{H}}(n,m) = 2^{nm}\}$

The following, which appears as corollary 13 in [1],[2] is a bound on the generalization error in terms of $d_{\mathbf{H}}$.

Theorem 3.8. *[1] Let* $\mathbf{H}$ *be any permissible boolean hypothesis space family.*[3] *If the number of examples* m *of each task satisfies*

$$m \geq \frac{88}{\epsilon^2}\left[2d_{\mathbf{H}}(n)\log\frac{22}{\epsilon} + \frac{1}{n}\log\frac{4}{\delta}\right],$$

then with probability at least $1-\delta$ *(over the choice of* $D_1, \ldots, D_n$*), for any* $H \in \mathbf{H}$*, and* $h_1, \ldots, h_n \in H$*,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i)\right| \leq \epsilon.$$

4. GENERALIZATION BOUNDS FOR THE AGNOSTIC SETTING

Let $\mathcal{X}$ be our domain set, let $\mathbb{H} \subseteq 2^{\mathcal{X}}$ denote the hypothesis class that we work with, and let $\mathcal{F}$ be the set of potential semantic transformations. Formally, we assume that $\mathcal{F}$ is a set of functions $f : \mathcal{X} \to \mathcal{X}$ such that $\mathcal{F}$ is a group under function composition and such that $\mathbb{H}$ is closed under the action of $\mathcal{F}$, i.e., for all $h \in \mathbb{H}$ and $f \in \mathcal{F}$, there exists $h' \in \mathbb{H}$ such that $x \in h \Leftrightarrow f(x) \in h'$.

We define equivalence relation $\sim_{\mathcal{F}}$ on $\mathbb{H}$ by

$$h \sim_{\mathcal{F}} h' \text{ iff } \exists f \in \mathcal{F} \text{ such that } h' = h \circ f$$

Let P be a distribution on $\mathcal{X} \times \{0,1\}$, and suppose we are given data sets $D_1, \ldots, D_n$, each containing at least m training examples, where examples $(y,b) \in D_i$ are generated by drawing (x,b) according to P, and applying transformation $f_i \in \mathcal{F}$ to x to obtain $y = f_i(x)$.

Our problem fits into the multitask learning setting as follows: The hypothesis space family, $\mathbf{H}$, is the family of all equivalence classes of $\sim_{\mathcal{F}}$, i.e. $\mathbf{H} = \mathbb{H}/\sim_{\mathcal{F}}$, and P_i is the distribution obtained by first generating (x,b) according to P, and outputting $(f_i(x), b)$.

$\mathcal{X}$	Domain of database entries (excluding the classifications)
$\mathcal{F}$	Family of possible transformations from $\mathcal{X}$ to $\mathcal{X}$, mapping between the different database schemas
P	A distribution on $\mathcal{X} \times \{0,1\}$
D_i	ith database, consisting of classified database entries, $(x,b) \in \mathcal{X} \times \{0,1\}$
n	The number of databases
m	The number of entries per database
ϵ	Desired prediction accuracy
δ	Tolerated probability of desired accuracy not being achieved
$\mathbb{H}$	Hypothesis space of possible classification functions from $\mathcal{X}$ to $\{0,1\}$
$\sim_{\mathcal{F}}$	Equivalence relation defined on $\mathbb{H}$: $h \sim_{\mathcal{F}} h'$ iff there is some $f \in \mathcal{F}$ mapping h to h'
$\mathbb{H}_{\sim_{\mathcal{F}}}$	Set of equivalence classes of $\mathbb{H}$ under $\sim_{\mathcal{F}}$
$d_{\mathbf{H}}(n)$	Generalized VC-dimension for collection $\mathbf{H}$ of hypothesis spaces

Table 1: Table of Symbols

Notation. Let $\mathbb{H}_{\sim_{\mathcal{F}}}$ denote $\mathbb{H}/\sim_{\mathcal{F}}$.

Explicitly, $\Pi_{\mathbb{H}_{\sim_{\mathcal{F}}}}(m,n) =$

$$\max_{\overline{x}_1,\ldots,\overline{x}_n \in X^m} \left| \left\{ \begin{bmatrix} \overline{h \circ f_1}(\overline{x_1}) \\ \vdots \\ \overline{h \circ f_n}(\overline{x_n}) \end{bmatrix} : f_1 \ldots f_n \in \mathcal{F}, h \in \mathbb{H} \right\} \right|$$

and $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n)$ is the maximal value of m for which this quantity is 2^{mn}.

Theorem 3.8 indicates that the generalization that may be guaranteed for such a learning setting is controlled by the value of this last parameter. In the following section we analyze $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n)$ as a function of the maximal VC-dimension of any equivalence class of $\mathbb{H}$ and the relationship between this hypothesis space and family $\mathcal{F}$.

4.1 Computation of $\mathbf{d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n)}$

We begin a simple upper bound on $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n)$.

Lemma 4.1. *For any* n, $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n) \leq$ *VC-dim*$(\mathbb{H})$.

PROOF. If there exist $\overline{x_1}, \ldots, \overline{x_n} \in \mathcal{X}$ such that

$$\left| \left\{ \begin{bmatrix} \overline{h_1}(\overline{x_1}) \\ \vdots \\ \overline{h_n}(\overline{x_n}) \end{bmatrix} : h_1 \sim_{\mathcal{F}} h_2 \ldots \sim_{\mathcal{F}} h_n \right\} \right| = 2^{mn},$$

then $\mathbb{H}$ shatters $\overline{x_i}$ for any $1 \leq i \leq n$. □

We will now prove several more interesting bounds on $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n)$. We first consider the case where $|\mathcal{F}|$ is finite.

Theorem 4.2. *If* $\mathcal{F}$ *is finite and* $\frac{n}{\log(n)} \geq$ *VC-dim*$(\mathbb{H})$*, then*

$$d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n) \leq 2\log(|\mathcal{F}|)$$

[2] Note that although [1] only states that $\frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) \leq \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i) + \epsilon$, it is clear from the proofs in [1] that this stronger form holds.

[3] Permissibility [1, 2] is a "weak measure-theoretic condition satisfied by almost all 'real-world' hypothesis space families". Throughout this paper we shall assume that all our classes are permissible.

PROOF. For fixed $\overline{x_1}, \ldots \overline{x_n} \in \mathcal{X}^m$, we have to bound the number of distinct matrices of the form

$$\begin{bmatrix} \overline{h \circ f_1}(\overline{x_1}) \\ \vdots \\ \overline{h \circ f_n}(\overline{x_n}) \end{bmatrix}$$

for $f_1, \ldots, f_n \in \mathcal{F}$, $h \in \mathbf{H}$.

Such a matrix is obtained by applying h to the vector $(\overline{f_1}(\overline{x_1}), \ldots, \overline{f_n}(\overline{x_n}))$.

Obviously, there are at most $|\mathcal{F}|^n$ many ways of choosing $f_1, \ldots, f_n$. Once this sequence of transformations is set, the matrix is determined by the dichotomy induced by the h on the above vector.

Let D denote VC-dim($\mathbb{H}$). Applying Sauer's Lemma, the number of dichotomies that may be induced by $h \in \mathbb{H}$ on a vector of $m \times n$ many points is at most $\left(\frac{enm}{D}\right)^D$.

It follows that the number of possible matrices is upper bounded by $\left(\frac{enm}{D}\right)^D \times |\mathcal{F}|^n$. The proof is now concluded by noting that for $m \geq 2\log(|\mathcal{F}|)$, it is the case that $\left(\frac{enm}{D}\right)^D \times |\mathcal{F}|^n \leq 2^{mn}$ □

We now consider $\mathcal{F}$ such that $|\mathbb{H}/\sim_{\mathcal{F}}|$ is finite.

Theorem 4.3. *If $\sim_{\mathcal{F}}$ is of finite index k, and $n \geq \frac{\log k}{4d \log d}$, then*

$$d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n) \leq \frac{\log k}{n} + 4d \log d,$$

for $d = \max\left\{3, \max_{H \in \mathbb{H}/\sim_{\mathcal{F}}} VC\text{-}dim(H)\right\}$.

PROOF. Fix $h \in \mathbb{H}$ and let $H = [h]_{\sim_{\mathcal{F}}}$. Since, by Sauer's lemma, $\Pi_H(m) \leq \left(\frac{em}{d}\right)^d \leq m^d$, H contributes at most m^d distinct rows for $n \times m$ matrices to be counted in $\Pi_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n, m)$. So, H contributes at most m^{nd} to this quantity, and thus $\Pi_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n, m) \leq km^{nd}$.

Now, if $\Pi_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n, m) \geq 2^{mn}$, then $km^{nd} \geq 2^{mn}$. Equivalently, $\frac{\log k}{n} + d \log m \geq m$.

So, suppose $n \geq \frac{log k}{4d \log d}$, and consider $m_0 = \frac{log k}{n} + 4d \log d$. Observe that

$$\frac{\log k}{n} + d \log m_0$$

$$= \frac{\log k}{n} + d \log\left(\frac{\log k}{n} + d \log d\right)$$

$$\leq \frac{\log k}{n} + d \log(4d \log d + 4d \log d)$$

$$= \frac{\log k}{n} + d \log(8d \log d) < m_0.$$

By monotonicity, for $n \geq \frac{log k}{4d \log d}$ and $d \geq 3$, $d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n) \leq m_0$. □

Finally, we proceed to generalize this last theorem to $\mathcal{F}$ such that $\sim_{\mathcal{F}}$ is of infinite index.

Definition 4.4. *For $h \in \mathbb{H}$, and $\overline{x_1}, \ldots, \overline{x_n} \in \mathcal{X}^m$ define* $splits_h(\overline{x_1}, \ldots, \overline{x_n})$

$$= \left\{ \begin{bmatrix} \overline{h_1}(\overline{x_1}) \\ \vdots \\ \overline{h_n}(\overline{x_n}) \end{bmatrix} : h_1, \ldots, h_n \in [h] \right\}$$

Definition 4.5. *For fixed $\overline{x_1}, \ldots, \overline{x_n} \in \mathcal{X}^m$, define*

$$h \leq_{(\overline{x_1}, \ldots, \overline{x_n})} h'$$

iff

$$splits_h(\overline{x_1}, \ldots, \overline{x_n}) \subseteq splits_{h'}(\overline{x_1}, \ldots, \overline{x_n})$$

Definition 4.6. *Let*

$$H_{max}(n) = \{h : h \text{ is } \leq_{(\overline{x_1}, \ldots, \overline{x_n})} \text{-maximal}\}.$$

For $h, h' \in H_{max}(n)$, define

$$h \sim_{(\overline{x_1}, \ldots, \overline{x_n})} h'$$

iff

$$splits_h(\overline{x_1}, \ldots, \overline{x_n}) = splits_{h'}(\overline{x_1}, \ldots, \overline{x_n}).$$

Definition 4.7. *Define* $index_{(\mathcal{F},m,n)}(\mathbb{H}) =$

$$\sup_{\overline{x_1}, \ldots, \overline{x_n} \in \mathcal{X}^m} (index \text{ of } \sim_{(\overline{x_1}, \ldots, \overline{x_n})} \text{ on } H_{max}(n)).$$

Finally, let $D = VC\text{-}dim(\mathbb{H})$, and define

$$index_{(\mathcal{F},n)}(\mathbb{H}) = \sup_{1 \leq m \leq D} index_{(\mathcal{F},m,n)}(\mathbb{H})$$

Theorem 4.8. *If $index_{(\mathcal{F},n)}(\mathbb{H}) = \phi(n)$, and*

$$n \geq \frac{\log \phi(n)}{4d \log d},$$

then

$$d_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n) \leq \frac{\log \phi(n)}{n} + 4d \log d,$$

where $d = \max\left\{3, \max_{H \in \mathbb{H}/\sim_{\mathcal{F}}} VC\text{-}dim(H)\right\}$.

PROOF. Recall that in the proof of theorem 4.3, we counted a contribution from each $H = [h]_{\sim_{\mathcal{F}}}$ of $\left(\frac{em}{d}\right)^{nd}$ towards $\Pi_{\mathbb{H}_{\sim_{\mathcal{F}}}}(n, m)$. Observe in the first paragraph of the proof that if $splits_{(\mathcal{F},\overline{x_1},\ldots,\overline{x_n})}(h) \subseteq splits_{(\mathcal{F},\overline{x_1},\ldots,\overline{x_n}))}(h')$, then any contribution counted from H in this argument would also be contributed by $H' = [h']_{\sim_{\mathcal{F}}}$. Thus, we only need count these contributions from $index_{(\mathcal{F},m,n)}(\mathbb{H})$ many $\sim_{\mathcal{F}}$-equivalence classes. The theorem follows. □

4.2 Main Results

We are now ready to state our main results. As before, we have distribution P over $\mathcal{X} \times \{0, 1\}$, family $\mathcal{F}$ of transformations on $\mathcal{X}$, closed under inverse and composition, hypothesis space $\mathbb{H}$ closed under the action of $\mathcal{F}$, and data sets $D_1, \ldots, D_n$, such that each D_i consists of at least m examples $(f_i(x), b)$, for some fixed $f_i \in \mathcal{F}$, where each (x, b) is drawn independently according to P.

Theorem 4.9. *For any $0 \leq \epsilon, \delta \leq 1$, and any $\sim_{\mathcal{F}}$-equivalent $h_1, \ldots, h_n \in \mathbb{H}$, if any of the following conditions hold, and D_i are generated randomly as described above, then with probability at least $1 - \delta$,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} Er^{P_i}(h_i) - \frac{1}{n} \sum_{i=1}^{n} \hat{Er}^{D_i}(h_i) \right| \leq \epsilon.$$

1. *$\mathcal{F}$ is finite,*

$$\frac{n}{\log(n)} \geq \max_{H \in \mathbb{H}/\sim_{\mathcal{F}}} (VC\text{-}dim(H))$$

and

$$m \geq \frac{88}{\epsilon^2}\left[4\left(\log\frac{22}{\epsilon}\right)\log(|\mathcal{F}|) + \frac{1}{n}\log\frac{4}{\delta}\right]$$

2. $\sim_{\mathcal{F}}$ *has finite index* k *on* $\mathbb{H}$, $n \geq \frac{\log k}{4d \log d}$, *and*

$$m \geq \frac{88}{\epsilon^2}\left[2\left(\log\frac{22}{\epsilon}\right)\left(\frac{\log k}{n} + 4d\log d\right) + \frac{1}{n}\log\frac{4}{\delta}\right],$$

where $d = \max\left\{3,\ \max_{H\in\mathbb{H}/\sim_{\mathcal{F}}}(VC\text{-}dim(H))\right\}$

3. $index_{(\mathcal{F},n)}(\mathbb{H}) = \phi(n)$, $n \geq \frac{\log \phi(n)}{4d \log d}$, *and*

$$m \geq \frac{88}{\epsilon^2}\left[2\left(\log\frac{22}{\epsilon}\right)\left(\frac{\log \phi(n)}{n} + 4d\log d\right) + \frac{1}{n}\log\frac{4}{\delta}\right],$$

where $d = \max\left\{3,\ \max_{H\in\mathbb{H}/\sim_{\mathcal{F}}}(VC\text{-}dim(H))\right\}$.

This theorem gives sufficient conditions on the size of the data sets to guarantee that (with high probability) the equivalence class with optimal average empirical error on the data sets contains a hypothesis which is which is near-optimal on average over the data sets for the underlying distribution. This is a first step[4] in justifying the following meta-algorithm:

1. Find an equivalence class $[h]_{\sim_{\mathcal{F}}}$ of hypotheses that minimizes
$$\inf_{h_1,\ldots,h_n\in[h]_{\sim_{\mathcal{F}}}} \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i).$$
2. Use standard learning techniques on D_j to select $h' \in [h]_{\sim_{\mathcal{F}}}$ which performs well on D_j.

Part 1 of theorem 4.9 says that for finite sets of transformations, once the number of data sets is sufficiently large, the logarithm of the size of the set of potential transformations may replace the VC-dimension of the hypothesis class $\mathbb{H}$ for the purpose of deriving sample complexity guarantees. Parts 2 and 3 say that this VC-dimension may be replaced by a function of the maximal VC-dimension of any of the *equivalence classes* of the hypothesis class and a measure of the complexity of the interaction between $\mathbb{H}$ and $\mathcal{F}$.

Finally, we present part 3 of theorem 4.9 as a bound on the generalization error in terms of the other parameters. (We omit the analogous statements for parts 1 and 2.)

Corollary 4.10. *Let* $0 \leq \epsilon, \delta \leq 1$, *let*

$$d = \max\left(\max_{H\in\mathbb{H}/\sim_{\mathcal{F}}}(VC\text{-}dim(H)), 3\right),$$

and let $index_{(\mathcal{F},n)}(\mathbb{H}) = \phi(n)$, *where* $D = VC\text{-}dim(\mathbb{H})$. *If* $n \geq \frac{\log \phi(n)}{8d \log d}$, $m \geq 3$, *and*

$$\epsilon \geq \frac{12}{m}\sqrt{2\left(\frac{\log \phi(n)}{n} + 8d\log d\right)\log m + \frac{1}{n}\log\frac{4}{\delta}},$$

with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i)\right| \leq \epsilon.$$

One drawback of the above results is that they all bound the *average*, over all n data sets, discrepancy between our empirical estimates and the true error of hypotheses. In a subsequent paper, [12], we have been able to improve these results to bounds that apply to any selected single data set. Namely, the current bounds that guarantee

$$\left|\frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i)\right| \leq \epsilon$$

can be shown, with no extra cost in sample size, to guarantee for every $j \leq n$,

$$\left|\inf_{h'\in[h]_{\sim_{\mathcal{F}}}} Er^{P_j}(h') - \inf_{h_1,\ldots,h_n\in[h]_{\sim_{\mathcal{F}}}} \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{D_i}(h_i)\right| \leq \epsilon.$$

[4]The second step is to prove that the average true error is close to the true error on any one of the data sets. We have recently succeeded in proving a result [12] along these lines.

5. RELATED WORK

Previous work on database integration has focused on schema matching, the problem of producing semantic mappings which transform data instances from one schema to instances of another. ([11] provides a comprehensive survey of schema matching.) This has been approached primarily by considering the schemas involved and using linguistic information (e.g. attribute labels which are synonyms or homonyms) and/or constraint information, such as data types, value ranges, and cardinalities, which is usually included in schemas [10, 4, 9]. However, some recent work has made use of both schema information and machine learning on actual data [5, 6, 17].

Our approach differs from this previous work in two significant ways. First, it focuses on the particular task of data integration for classification prediction, as opposed to complete unification of multiple databases into a single database. Our work suggests that it may be advantageous to focus on data integration with a particular use of the data in mind, as the available data may be sufficient for this use but insufficient for determining the semantic mappings.

The other important difference between our work and prior database integration work is that our methods make no use of the database schemas, which is advantageous as the schema information may be incomplete or inaccurate.

We should point out the distinction between our notion of "data integration" and the concept of "data fusion," on which there is an abundance of literature [7, 15, 16]. Data fusion seeks to integrate data from sources that are disparate in a much stronger sense than the one we have considered here. Whereas we have assumed that the different sources are merely transformed versions of one another, in data fusion, each of the sources actually provide *different kinds of information* about some common phenomenon. The different sources may differ significantly not only in the representation of the data, but also in the type of information and even the accuracy of the information. A typical example involves networks of different types of sensors in different locations. This is clearly a more difficult problem than the one we have considered here.

Another relevant line of research involves learning multiple "related" tasks (multitask learning) [1] and "learning to learn" [14]. The former addresses the problem of simultaneously learning multiple related tasks, while the latter considers learners as embedded in an environment in which learning of prior tasks improves the ability of the learner to learn future tasks. While these approaches have not previously been applied to data integration, they are well-suited to this problem, as the tasks of learning different schemas for similar databases are indeed tightly related tasks, for which it is natural to expect to gain useful knowledge from each

task. Furthermore, we feel that our work sheds some light on what it means for tasks to be "related," a concept whose formal definition has been sufficiently elusive to all but halt progress on the theoretical side of multitask learning (in spite of the plethora of promising experimental work in this area, e.g., [3, 13]). Perhaps our concrete concept of "related" and successful analysis of multitask learning within this framework will inspire further formal development in this area.

6. CONCLUSIONS

In this work we have attempted to provide a formalism to support the application of the mathematical machinery of computational learning theory to the task of classification prediction utilizing disparate data sources. The results we obtain justify the use of an empirical risk minimization approach in this domain. In particular they show that data sets consisting of transformed training samples may be utilized on the basis of knowing a set of potential data transformations even when the actual semantic mappings applied to each data set are not known.

The analysis shows that the information complexity of such an approach depends on the richness of both the hypothesis class used for prediction (or regression) and the set of potential data transformations. Furthermore, this sample complexity also depends on the relationship between these two families of functions. The crucial factor in determining this sample complexity turns out to be the equivalence relation induced on the hypothesis class by the set of data transformations. This is formally measured by the VC-dimension of the resulting equivalence classes of hypotheses and by the index of this equivalence relation.

The strongest results in this work are obtained for the specific case of learning rectangles, in the PAC setting, under the set of multidimensional data shifts. In that case we prove that, for the purpose of learning the shape of the rectangles, disparate data sets are as effective as a single training data set of size close to the size of their union.

We view this work as a kind of appetizer for further applications of COLT techniques to issues of obtaining collective knowledge from a set of varied data sources. This work leaves many loose ends hanging, and we hope they will stimulate further research.

7. ACKNOWLEDGMENTS

We wish to thank Matthew Schultz for raising the questions that initiated this research, and to Rich Caruana for helpful discussions concerning this work.

8. REFERENCES

[1] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

[3] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[4] S. Castano and V. D. Antonellis. A schema analysis and reconciliation tool environment for heterogeneous databases. In *IDEAS*, pages 53–62, 1999.

[5] A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.

[6] R. Goldman and J. Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 436–445. Morgan Kaufmann, 1997.

[7] D. Hall and J. Llinas. An introduction to multisensor data fusion. In *Proceedings of the IEEE*, volume 85, pages 6 –23.

[8] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[9] D. S. Luigi Palopoli and D. Ursino. Semi-automatic semantic discovery of properties from database schemas. In *IDEAS*, pages 244–253, 1998.

[10] T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 122–133, 24–27 1998.

[11] E. Rahm and P. Bernstein. On matching schemas automatically. *Dept. of Computer Science, Univ. of Leipzig*, 2001.

[12] S. Ben-David, J. Gehrke and R. Schuller. Technical report, Computer Science Department, Cornell University, May 2002.

[13] S. Thrun and J. O'Sullivan. Discovering structure in multiple learning tasks: The TC algorithm. In *International Conference on Machine Learning*, pages 489–497, 1996.

[14] S. Thrun and L. Pratt, editors. *Learning To Learn*. Kluwer Academic Publishers, November 1997.

[15] H. Wache, T. Vgele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hbner. Ontology-based integration of information - a survey of existing approaches. In *Proceedings of the Workshop Ontologies and Information Sharing, IJCAI*, 2001.

[16] L. Wald. An overview of concepts in fusion of earth data. In P. Gudmandsen, editor, *Future trends in Remote Sensing*, pages 385 – 390. Balkema, 1997.

[17] Q. Y. Wang, J. X. Yu, and K.-F. Wong. Approximate graph schema extraction for semi-structured data. In *Advances in Database Technology - EDBT 2000, 7th International Conference on Extending Database Technology, Konstanz, Germany, March 27-31, 2000, Proceedings*, volume 1777 of *Lecture Notes in Computer Science*, pages 302–316. Springer, 2000.