Auto-correlation Dependent Bounds for Relational Data

Amit Dhurandhar
adhuran@us.ibm.com
Mathematical Sciences Dept.
IBM T.J. Watson
1101 Kitchawan Road
Yorktown Heights, USA

ABSTRACT

A large portion of the data that is collected in various application domains such as online social networking, finance, biomedicine, etc. is relational in nature. A subfield of Machine Learning namely; Statistical Relational Learning (SRL) is concerned with performing statistical inference on relational data. A defining property of relational data that separates it from independently and identically distributed data (i.i.d.) is the existence of correlations between individual datapoints. A major portion of the theory developed in machine learning assumes the data is i.i.d. In this paper we develop theory for the relational setting. In particular, we derive distribution free bounds for the relational setting where the class of data generation models we consider are inspired from the type joint distributions that are represented by relational classification models developed by the SRL community. A key aspect of the bound we derive is that the tightness of the bound is a function of the strength of dependence between related datapoints, with the bound reducing to the standard Hoeffding's or McDiarmid's inequality when there is no dependence. To the best of our knowledge this is the first bound for relational data whose tightness varies with the strength of dependence.

1. INTRODUCTION

Traditional Machine Learning primararily considers modeling of independently and identically distributed (i.i.d.) data. However, real life data is rarely i.i.d. with correlations existing between various datapoints. Such non-i.i.d. or relational data occurs in various domains ranging from biology to finance. A new emerging sub-area of Machine Learning namely; Statistical Relational Learning (SRL) [11] is concerned with modeling of uncertainty in such type of non-i.i.d. or relational data.

Collective classification is one of the important problems considered in SRL. In collective classification related data instances are classified simultaneously rather than independently as done in traditional classification. Though there are numerous relational classification algorithms [27, 29, 10, 22] developed in literature, the current state of theory – distribution free bounds, for relational domains in general is still primitive when compared with the traditional setting. The need for developing such theory has been expressed in [14, 11].

Distribution free bounds derived in Machine Learning, are used to bound the empirical error (i.e. test or training error) of a classifier with respect to (w.r.t.) its generalization error. The generalization error of a classifier is the expected error of a classifier over the entire input w.r.t. the underlying distribution. Hence, the generalization error is also referred to as the true error. If we are to evaluate a particular classifier or choose the best classifier amongst available options, the generalization error can serve as a great yardstick. Unfortunately, this error cannot be computed directly, since the true underlying distribution of the sample is unknown. The empirical error on the other hand can be computed from the sample. Distribution free bounds relate these two errors by providing us with probabilistic estimates for the generalization error given the empirical error without knowledge of the underlying distribution. This is the main advantage of having these bounds.

Various distribution free bounds have been derived in Statistics and Machine Learning literature. The Markov inequality [24, 12], the Chebyshev inequality [24, 12] and the Hoeffding inequality [13] which bound a random variable to its mean are amongst the most popular. The Hoeffding inequality however, gives tighter bounds than these two inequalities when the sample size increases [13]. Other such inequalities are given by Chernoff [5], Bennett [3] and Okamoto [23]. Distribution free bounds on the generalization error of a classifier are provided by Vapnik [30] based on a property of the classifier space called Vapnik-Chervonenkis (VC) dimension. In [8] distribution free bounds are provided for the k-nearest neighbor algorithm. In [20] improved Probably Approximately Correct (PAC) Bayes bounds are provided for linear decoders. These bounds are tighter than the ones previously introduced in [19]. In [4] bounds are provided for a validation technique called progressive validation which are tighter than those for hold-out-set validation. The derivation of these bounds uses Hoeffdings inequality thus portraying its widespread use in Machine Learning. A nice survey explaining the pros and cons of these different bounds used in Machine learning is given in [18]. One of the main conclusions of this survey is that test set bounds (i.e. empirical error is the test error) are generally tighter and easier to apply than training set bounds. In this paper we derive

Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA

a test set bound for relational data which will be different from the bounds we have discussed so far since they all apply to i.i.d. data.

Bounds for non-i.i.d. data have been derived in specific settings. Some of the well known settings where such bounds have been derived are in time series analysis and pseudo-random number generation. In time series analysis, data is assumed to come sequentially in time from an underlying data generation process. While deriving bounds in this setting the main assumptions on the data generation process are that it is stationary in time and the strength of dependence between datapoints decays as they are more separated in time (β mixing or ϕ mixing processes) [17, 21]. Stationary in this case means that any k consecutive datapoints chosen from this stream of data have the same distribution. This setting is very different from the relational setting we consider since in our setting we do not make the above two assumptions. In pseudo-random number generation a limited notion of independence is assumed which is called k-wise independence. In k-wise independence any set of k (or fewer) random variables are assumed to be independent from a total of n random variables ($k \leq n$). Bounds assuming k-wise independence are given in [28] which extend the ideas given by Chernoff and Hoeffding. This setting is also significantly different from ours since the assumption of k independence is unrealistic for the applications we mentioned before.

A number of learnability results (both positive and negative) have been proven for restricted classes of inductive logic programs [6, 25, 2]. The learnability results are primarily based on two formal models of learning namely; PAC learning and learning from equivalence and membership queries. Bounds for relational data in applications we are interested in have been derived in [9, 15, 26]. However, both of these derived bounds are indifferent to the strength of dependence between interacting datapoints. In other words, the bounds remain the same irrespective of how strongly correlated the interacting datapoints are. In this paper, we derive a bound that varies with the degree of dependence between related datapoints; which is an attractive feature. We hence, refer to our bound as strength of dependence bound (STB). Moreover, as we will see later the bound becomes the standard Hoeffding or McDiarmids inequality when the datapoints are not dependent.

The rest of the paper is organized as follows. In Section 2, we describe and motivate the data generation models we consider in this paper. In Section 3, we clearly state and justify the assumptions needed in obtaining the results. In Section 4, we state these results and then prove the main result in section 5. In Section 6, we show that our bound is robust to deviations from our primary assumption. We also compare our bound to other bounds in terms of tightness by applying it to real relational datasets. We discuss applicability of the results and suggest future lines of research in section 7.

2. PRELIMINARIES

Say N datapoints $(x_1, y_1, ..., x_N, y_N) \in (X \times Y)^N$ where X is the input space and Y is the output space are drawn from the joint distribution $P[X_1, Y_1, ..., X_N, Y_N]$. Note that $X_i \times Y_i \ \forall i \in \{1, ..., N\}$ denotes the i^{th} copy of the $X \times Y$ space. If the datapoints are i.i.d. the joint probability would factorize as follows: $P[X_1, Y_1, ..., X_N, Y_N] = P^N[X, Y]$. However, in the case of relational data certain dependencies may exist between datapoints which prevents this factorization. Hence, at one end of the spectrum we have dependencies between all N datapoints with the joint probability or the underlying distribution having the following form $P[X_1, Y_1, ..., X_N, Y_N]$, whereas at the other end of the spectrum all the N datapoints are i.i.d. with the underlying distribution being specified over the $X \times Y$ space having the form $P[X, Y]$. There are a range of distributions that lie between these two extremeties where the dependence is amongst disjoint subsets of datapoints with independence between these subsets. For example, given N datapoints the first m_1 may be related and then the next m_2 may be related ($m_1 + m_2 = N$) with independence between these two subsets. In this case the underlying distribution would have the following form, $P[X_1, Y_1, ..., X_N, Y_N] = P[X_1, Y_1, ..., X_{m_1}, Y_{m_1}]P[X_{m_1+1}, Y_{m_1+1}, ..., X_N, Y_N]$. The distributions over the two subsets may not be the same but they are independent with their product being the probability of the given dataset. We can have many such distributions with different number of subsets, different sizes of the subsets (summing to N) and different datapoints being involved in each subset. In Section 4 we will derive distribution free bounds that apply to this entire spectrum of data generation models. We will now define certain basic concepts and motivate the above data generation models by showing that the state-of-the-art relational classification models actually represent joint distributions that have this form.

Relational data: Relational data consists of objects and the relationships between these objects are termed as links. Each object and link have a *type* associated with them. Objects or links of the same *type* have the same set of attributes. Relational data can be represented at the *type* level by a graph which is called a relational schema whereas relational data represented at the individual object and link level as a graph is called a relational data graph (or instance graph) [22], wherein the vertices are the objects and the edges are the links. An example relational schema and the corresponding data graph (i.e. the actual dataset) are shown in Figures 1a and 1b respectively. The relational schema has 2 object types namely; *Paper* and *Author*. The data graph shows 2 authors linked to the papers they authored or co-authored.

Probabilistic Models over Relational Data: Probabilistic Models over relational data (PMRD) [11] are structured graphical models that are used to handle uncertainity in relational domains. These models include but are not limited to probabilistic relational models, relational dependency networks, different markov networks. A PMRD represents a joint distribution over the attributes of a data graph. Consider Figure 1a where the object type *Paper* has 2 attributes, Title and Area which imply the title of the paper and the research area it belongs to respectively. Let the attribute Area be the class label i.e. we want to classify papers based on their research area. The object type *Author* has attributes Paper Title and Age, which relates a particular paper to the ages of the authors that wrote it. The Title attribute (a primary key) in *Paper* is the same as the Paper Title attribute (a foreign key) in *Author*. Hence, each Paper object has 3 attributes namely; Title, Area and Age. The attributes Title and Area are called *intrinsic attributes* as they belong to object type *Paper* and the attribute Age is called a *relational attribute* since it belongs to a different linked object type *Author*. Each paper can have variable number of authors and thus each paper would be associated with

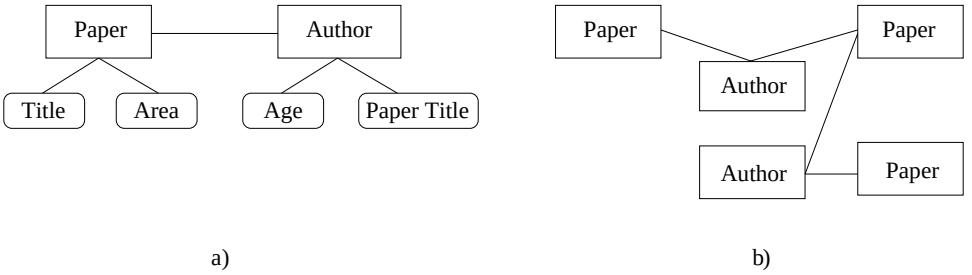

Figure 1: a) represents a relational schema with object types, *Paper* and *Author*. The relationship between them is many-to-many. The rounded boxes linked to these object types denote their respective attributes. b) is the corresponding data graph which shows authors linked to the papers that they authored or co-authored.

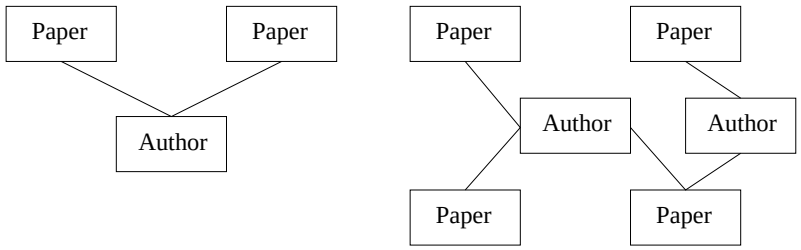

Figure 2: Above we see a disconnected data graph with 2 components of variable size (i.e. size 2 component on the left and size 4 component on the right).

multiple values of Age. A popular solution to this problem is to aggregate the values of the attribute Age of *Author* into a single value such that each paper is associated with only a single Age value. An aggregation function such as average over the ages of the related authors for each paper can be used. Now instead of the Age attribute we can introduce a new attribute AvgAge which denotes average age. With this the attributes of Paper object are; Title, Area and AvgAge. Hence, the joint distribution represented by a PMRD on the data graph in Figure 1b is,

$$P[A_1, A_2, A_3]$$

where A_i denotes the attribute set $\{Title^i, Area^i, AvgAge^i\}$ of the i^{th} Paper object. Since in Figure 1b we have paths connecting the 3 papers (through authors), we have 3 copies of the same attributes (which may have different values) in the joint distribution.

The data graph in Figure 1b is connected. It is possible in some other case that the data graph is actually disconnected. This is shown in Figure 2. The joint distribution over the data graph in Figure 2 is,

$$P[A_1, A_2, ..., A_6] = P[A_1, A_2]P[A_3, A_4, A_5, A_6]$$

since the data graph has 6 Paper objects, we have 6 copies of the attributes. Moreover, there are 2 disconnected components in the graph, one with 2 Paper objects and the other with 4 Paper objects. Consequently, the joint probability distribution $P[A_1, A_2, ..., A_6]$ factorizes as a product of 2 independent distributions $P[A_1, A_2]$ and $P[A_3, A_4, A_5, A_6]$.

Note that since these PMRDs learn at the template level the marginals over individual datapoints are identical to each other. In other words, given a distribution $P[A_1, A_2, ..., A_N]$ over a dataset of size N, $P[A_i] = P[A_j]\ \forall i, j \in \{1, ..., N\}$ irrespective of how the distribution factorizes. Realize that the above statement does *not* imply that the data is i.i.d. since the various A_i may depend on each other which prevents the i.i.d. factorization of the joint probability.

We have thus seen that the type of distributions/data generation models that we are going to derive bounds for, subsume the distributions represented by these PMRDs that are extensively used to model relational data in practice.

Generalization Error (GE): Let $P[X, Y]$ be a distribution over the input-output space. A classifier $\zeta(.)$ takes as input a particular $x \in X$ and outputs a particular class label $y \in Y$. Let $\lambda(.,.)$ denote a bounded loss function outputing values in the range $[0, M]$ where M is a positive integer. Then the GE of ζ is defined as,

$$GE = E[\lambda(\zeta(x), y)]$$

In case of the relational setting the x maybe not just ones own attributes but in addition, attributes and class labels of related datapoints.

Hold-out Error (HE): The test error or the hold-out error (HE) computed over a test set of size N is given by,

$$HE = \frac{\sum_{i=1}^{N} \lambda(\zeta(x_i), y_i)}{N}$$

where $x_i \in X$, $y_i \in Y$ and y_i is the true label of x_i.

Strength of Dependence (d): In the case of relational data, we define d as the absolute value of relational autocorrelation ρ, which measures the degree of statistical dependence of the class label on related/linked datapoints [22]. A popular choice to measure ρ is (normalized) relative entropy [11]. The value of d lies in the interval $[0, 1]$, where 0 means the datapoints are uncorrelated while 1 means that the datapoints are highly correlated.

3. SETUP

In this section we set the stage for the next section where

the actual theoretical results are presented. The setup we describe here and the main technical result presented in the next section, is for a more general setting than our relational setting. We do this since it simplifies the proof (ignoring unnecessary details) and possibly allows the result to be used in a wider range of applications than those that are considered here.

Let m (exchangeable) random variables $Z_1, ..., Z_m$ be distributed according to the joint distribution $P[Z_1, ..., Z_m]$ such that no proper subset of these random variables is independent of the rest i.e. the joint probability cannot be written as a product of two or more independent distributions[1]. Since the joint probability can be factorized as: $P[Z_1, ..., Z_m] = P[Z_1]P[Z_2|Z_1]...P[Z_m|Z_{m-1}, ..., Z_1]$, we can view the points $z_1, ..., z_m$ as being sampled sequentially from distributions $P[Z_1]$, $P[Z_2|Z_1]$, $\cdots$, $P[Z_m|Z_{m-1}, ..., Z_1]$ respectively. *Note that we do not need to know the exact sampling order to apply the results in the next section.* Just the existence of some such ordering (which always does) suffices. Relating this back to the previous sections, we can view the Z_i as being any deterministic function defined over the input-output space.[2]

The two main assumptions we make that help us in deriving the results in the next section are as follows:

3.1 Assumption 1

To derive an inequality that depends on the strength of dependence d, we have to characterize how d affects the nature of the dependence between the random variables. What is probably desirable is that, as d tends to 0 the relationship we assume should look increasingly like the independent case and as d tends to 1 the relationship should reflect the high level of similarity to the previously sampled datapoints. We incorporate these ideas into our assumption in the following simple way,

$$\forall i \in \{2, ..., m\} \quad E[Z_i|Z_{i-1} = z_{i-1}, ..., Z_1 = z_1]$$
$$= d\frac{\sum_{k=1}^{i-1} z_k}{i-1} + (1-d)E[Z_i]$$

In the above equation d acts as a slider variable which controls the influence of the two terms on the right hand side. As d approaches 0 the conditional expectation depends less and less on the variables it is conditioned on and eventually takes the form of an unconditional expectation. On the other hand, as d approaches 1 the conditional expectation depends more heavily on the variables it is conditioned on.

From a relational setting point of view, the d is computed over the data and Z_i could be viewed as zero-one loss functions and we would expect any reasonable classification algorithm to give highly correlated errors if the datapoints are highly similar (i.e. large d) and uncorrelated errors if they are independent. This is precisely the intuition that the assumption captures.

Assumption 1 will hold exactly at the extremeties when d is zero and there is independence or when d is 1 and all the variables take on the same value giving rise to a martingale. The assumption is mainly just a simple way of incorporating our intuitions of how the relationship should look with varying d.

[1]The joint probability can also be seen as an independent component of a larger probability distribution over more variables.

[2]Since the input-output space is randomly generated, the Z_i are random variables.

3.2 Assumption 2

In the previous sections we have seen that PMRD's learn at the template level and hence, $P[A_1] = P[A_2] = ... = P[A_m]$ where A_i $i \in \{1, 2, ..., m\}$ denotes the relevant input-output space for the i^{th} instance. As mentioned before, Z_i is a deterministic function applied to the input-output space of the i^{th} instance i.e. to A_i. These two facts together imply, $\forall z \in \mathcal{Z}, P[Z_1 \leq z] = P[Z_2 \leq z] = ... = P[Z_m \leq z]$ where $\mathcal{Z}$ is the range of the random variables. This assumption was made in [26, 9]. Please note that this does not mean that the samples are i.i.d. The i.i.d. assumption would require independence between these random variables in addition to the distributions being the same. In this paper, however, we make the much less stringent assumption of just the means of the Z_i being equal, i.e.

$$E[Z_1] = E[Z_2] = ... = E[Z_m]$$

This implies that higher (> 1) moments of the Z_i are not required to be equal for proving the second theorem. Hence, interestingly, the weaker assumption of the expectations being equal suffices in this case.

4. RESULTS

In this section we first state the general result in Theorem 1 which only requires Assumption 1 to be true. We then show, how when the strength of dependence is zero, the inequality in Theorem 1 reduces to the well known Hoeffding inequality applicable to bounded independent random variables. We then customize the general result to our relational setting in Theorem 2 which requires both the assumptions in the previous section to be true.

We assume that our sample is of size N and the number of independent subsets is k, where T_i $i \in \{1, 2, ..., k\}$ represents the corresponding subset. m_i is the number datapoints in subset T_i. If we order the sample such that datapoints in T_i precede datapoints in T_j $\forall i < j \in \{1, 2, ..., k\}$ then $g(i)$ is the offset of the first datapoint in T_i i.e. 1 plus the sum of all m_j such that $j < i$ and assuming $m_0 = 0$.

THEOREM 1. *Let N points $(z_1, ..., z_N)$ be drawn sequentially from $P[Z_1, ..., Z_N] = \prod_{i=1}^{k} T_i$ where $k \in \{1, ..., N\}$ is the number of disjoint independent subsets of the random variables. Let T_i be a joint distribution over m_i consecutive attributes in $Z = (Z_1, Z_2, ..., Z_N)$ that are dependent such that $\sum_{l=1}^{k} m_l = N$ and if $i < j \in \{1, ..., k\}$ then the attribute with the highest index in T_i is strictly less than the attribute with the least index in T_j. For $i \in \{1, ..., N\}$, $g(r+1) > i > g(r)$ we assume $E[Z_i|Z_{i-1}, ..., Z_{g(r)}] = d_r \frac{\sum_{j=g(r)}^{i-1} z_j}{i-g(r)} + (1-d_r)E[Z_i]$ where $a_i \leq Z_i \leq b_i$, $g(r) = 1 + \sum_{j=1}^{r} m_{j-1}$ with $m_0 = 0$, $m_{k+1} = 1$, $r \in \{1, ..., k\}$, $d_r \in [0, 1]$ is the strength of dependence between attributes in T_r and $\delta = max_{i,j \in \{1,...,N\}}(b_i - a_j)$, then we have for $t > \frac{\sum_{j=1}^{k}(m_j-1)\delta d_j}{N}$,*

$$P[|\bar{Z} - E[\bar{Z}]| \geq t] \leq 2e^{\frac{-2(Nt - \sum_{j=1}^{k}(m_j-1)\delta d_j)^2}{\sum_{j=1}^{N}(b_j-a_j)^2}}$$

where $\bar{Z} = \sum_{i=1}^{N} \frac{z_i}{N}$.

Above we have an exponential bound that depends on the size of the sample (N), the sizes of the subsets of datapoints that are related (m_j where $j \in \{1, ..., k\}$), the autocorrelation between datapoints in each subset (d_j where $j \in \{1, ..., k\}$) and the ranges of Z_i ($[a_i, b_i]$). The inequality being applicable for $t > \frac{\sum_{j=1}^{k}(m_j - 1)\delta d_j}{N}$ implies that when the strength of dependence between related datapoints is high (i.e. d_j are close to 1) and the number of independent subsets is low (i.e. k is close to 1), the probability of the difference between $\bar{Z}$ and its expected value being "small" is practically 0 and hence the question of the upper bound being applicable (i.e. less than 1) is reasonable to ask only for larger values of t. Also note that the T_i being defined over consecutive random variables is not a constraint since non-consecutive random variables in a joint probability can always be made consecutive by reordering them, giving rise to the same distribution.

COROLLARY 1. *In Theorem 1 if the $d_j = 0\ \forall j \in \{1, ..., k\}$ (i.e. the datapoints are independent) then for $t > 0$ we have,*

$$P[|\bar{Z} - E[\bar{Z}]| \geq t] \leq 2e^{\frac{-2N^2t^2}{\sum_{j=1}^{N}(b_j - a_j)^2}}$$

which is the Hoeffding inequality.

The derived inequality in Theorem 1 thus has this nice property that it reduces to a well known inequality in the independent case. In case of relational data we usually have a single strength of dependence parameter d for the entire dataset and all the Z_i are the same loss function $\lambda(.,.) \in [0, M]$ ($M > 0$) applied to each $(x_i, y_i) \in X \times Y$ in the following manner, $Z_i = \lambda(\zeta(x_i), y_i)$. $\zeta(.)$ is a classifier that outputs a class label $y \in Y$.

THEOREM 2. *If we have relational data, then given a single strength of dependence parameter d, a loss function $\lambda(.,.) \in [0, M]$, k independent subsets and assuming $E[\lambda_1] = E[\lambda_2] = ... = E[\lambda_N]$, we have from the setup in Theorem 1 for $t > \frac{(N-k)Md}{N}$,*

$$P[|HE - GE| \geq t] \leq 2e^{\frac{-2(Nt-(N-k)Md)^2}{NM^2}}$$

where $\lambda_i = \lambda(\zeta(x_i), y_i)\ \forall i \in \{1, ..., N\}$.

PROOF. By Assumption 2 we have, $E[\bar{Z}] = E[\sum_{i=1}^{N} \frac{\lambda_i}{N}] = \frac{1}{N}\sum_{i=1}^{N} E[\lambda_i] = E[\lambda_j] = GE$ where $\lambda_i = \lambda(\zeta(x_i), y_i)$ and $i, j \in \{1, ..., N\}$. Substituting this result in Theorem 1 we get Theorem 2. □

In the case of relational data Assumption 1 says that for d close to 1 the (expected) performance of a classifier on a datapoint is very similar to its performance on related datapoints and for d close to 0 the performance is unrelated to the performance on these datapoints. The stronger version of Assumption 2 says that the probability of sampling the first datapoint (x_1, y_1) is the same irrespective of the marginal it is sampled from.

5. PROOF OF THEOREM 1

PROOF. Let N points $(f_1, ..., f_N)$ be drawn sequentially from $P[F_1, ..., F_N] = \prod_{i=1}^{k} T_i$ where $k \in \{1, ..., N\}$ is the number of disjoint independent subsets of the random variables. Without loss of generality, let T_i be a joint distribution over m_i consecutive attributes in $F = (F_1, F_2, ..., F_N)$ that are dependent such that $\sum_{i=1}^{k} m_i = N$ and if $i < j \in \{1, ..., k\}$ then the attribute with the highest index in T_i is strictly less than the attribute with the least index in T_j. For $i \in \{1, ..., N\}$, $g(r+1) > i > g(r)$ we assume $E[F_i|F_{i-1}, ..., F_{g(r)}] = d_r \frac{\sum_{j=g(r)}^{i-1} f_j}{i - g(r)} + (1 - d_r)E[F_i]$ where $a_i \leq F_i \leq b_i$, $g(r) = 1 + \sum_{j=1}^{r} m_{j-1}$ with $m_0 = 0$, $m_{k+1} = 1$, $r \in \{1, ..., k\}$, $d_r \in [0, 1]$ is the strength of dependence between attributes in T_r. Hence, $T_i = P[F_{g(i)}, ..., F_{g(i)+m_i-1}]$ and notice that every pair T_i, T_j is independent $\forall i, j \in \{1, ..., k\}, i \neq j$.

We will upper bound $P[|\bar{f} - E[\bar{f}]| \geq t]$ by upper bounding $P[\bar{f} - E[\bar{f}] \geq t]$ and $P[E[\bar{f}] - \bar{f} \geq t]$ which have the same upper bound and then applying union bound. Note that $\bar{f} = \sum_{i=1}^{N} \frac{f_i}{N}$ and t is strictly positive.

$$P[\bar{f} - E[\bar{f}] \geq t] = P[\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] \geq Nt]$$

Now, $I[Z \geq 0] \leq e^{hZ}$ where $I[.]$ is an indicator function, Z is a random variable and h is any positive real number (i.e. $h > 0$). Consequently,

$$\begin{aligned} & P[\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] \geq Nt] \\ & = E[I[\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] - Nt \geq 0]] \\ & \leq E[e^{h(\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] - Nt)}] \\ & = e^{-hNt} \prod_{r=1}^{k} E[\prod_{i=g(r)}^{g(r)+m_r-1} e^{h(f_i - E[f_i])}] \end{aligned} \tag{1}$$

The expectations $E[\prod_{i=g(r)}^{g(r)+m_r-1} e^{h(f_i - E[f_i])}]$ do *not* factorize as a product of expectations since the respective f_i are dependent. If we let $Q_i = e^{h(f_i - E[f_i])}$ we have,

$$\begin{aligned} & E[\prod_{i=g(r)}^{g(r)+m_r-1} Q_i] \\ & = E[Q_{g(r)} \cdot E[Q_{g(r)+1} ... E[Q_{g(r)+m_r-1}|Q_{g(r)+m_r-2}..., Q_{g(r)}] \\ & \quad ...|Q_{g(r)}]] \end{aligned}$$

If we are able to upper bound $E[Q_i|Q_{i-1}, ..., Q_{g(r)}]\ \forall i \in \{g(r)+1, ..., g(r)+m_r-1\}$ by some value w_i independent of $f_{i-1}, ..., f_{g(r)}$ and $E[Q_{g(r)}]$ by some other value u we would have,

$$E[\prod_{i=g(r)}^{g(r)+m_r-1} Q_i] \leq u \prod_{i=g(r)+1}^{g(r)+m_r-1} w_i \tag{2}$$

The above inequality could then be used to upper bound $P[\bar{f} - E[\bar{f}] \geq t]$. If Z is a random variable such that $a \leq Z \leq b$ and since e^{hZ} is a convex function, then by Jensen's inequality we have,

$$e^{hZ} \leq \frac{b-Z}{b-a}e^{ha} + \frac{Z-a}{b-a}e^{hb}$$

Using the above inequality for $\forall i \in \{g(r)+1, ..., g(r)+m_r-1\}$ we have,

$$\begin{aligned}
&E[Q_i|Q_{i-1}, ..., Q_{g(r)}] \\
&= E[e^{h(f_i - E[f_i])}|f_{i-1}, ..., f_{g(r)}] \\
&\leq e^{-hE[f_i]}(\frac{b_i - E[f_i|f_{i-1}, ..., f_{g(r)}]}{b_i - a_i}e^{ha_i} \\
&+ \frac{E[f_i|f_{i-1}, ..., f_{g(r)}] - a_i}{b_i - a_i}e^{hb_i}) \\
&= e^{v(h)}
\end{aligned}$$

where $v(h) = -hE[f_i] + ln(\frac{b_i - E_i}{b_i - a_i}e^{ha_i} + \frac{E_i - a_i}{b_i - a_i}e^{hb_i})$ and $E_i = E[f_i|f_{i-1}, ..., f_{g(r)}]$. We transform the function $v(h)$ to $v(h_i)$ where $h_i = h(b_i - a_i)$ and $s_i = \frac{E_i - a_i}{b_i - a_i}$. Hence, by assumption 1 we have $v(h_i) = -h_i s_i + \frac{h_i}{b_i - a_i}d_r(\frac{\sum_{j=g(r)}^{i-1} f_j}{i - g(r)} - E[f_i]) + ln(1 - s_i + s_i e_i^h)$. We now upper bound the function $v(h_i)$ which is the same as upper bounding $v(h)$ by using Taylors theorem. Thus we have $v(0) = 0$, $v'(0) = \frac{d_r}{b_i - a_i}(\frac{\sum_{j=g(r)}^{i-1} f_j}{i - g(r)} - E[f_i]) \leq d_r\frac{\delta}{b_i - a_i}$ where $\delta = max_{i,j \in \{1,...,N\}}(b_i - a_j)$. The inequality is an equality for $d_r = 0$. Hence, we upper bound the second derivative of $v(h_i)$ at 0 i.e. $v''(0) = s_i(1 - s_i) \leq \frac{1}{4}$. This is so since $s_i \in [0, 1]$. Hence, by Taylors theorem we have,

$$\begin{aligned}
v(h) = v(h_i) &\leq d_r h_i \frac{\delta}{b_i - a_i} + \frac{1}{8}h_i^2 \\
&= d_r \delta h + \frac{1}{8}h^2(b_i - a_i)^2
\end{aligned}$$

Hence from the above two equations and since e^z (where $z \in (-\infty, \infty)$) is a monotonic function we have $\forall i \in \{g(r)+1, ..., g(r)+m_r-1\}$,

$$E[Q_i|Q_{i-1}, ..., Q_{g(r)}] \leq e^{d_r \delta h + \frac{1}{8}h^2(b_i - a_i)^2} \quad (3)$$

Note that the right side in the above inequality is not a function of $f_{i-1}, ..., f_{g(r)}$ and hence the bound on the expectation of products will look like in equation 2. Similarly, we can now bound $E[Q_{g(r)}]$,

$$\begin{aligned}
E[Q_{g(r)}] &= E[e^{h(f_{g(r)} - E[f_{g(r)}])} \\
&\leq e^{-hE[f_{g(r)}]}(\frac{b_{g(r)} - E[f_{g(r)}]}{b_{g(r)} - a_{g(r)}}e^{ha_{g(r)}} \\
&+ \frac{E[f_{g(r)}] - a_{g(r)}}{b_{g(r)} - a_{g(r)}}e^{hb_{g(r)}}) \\
&= e^{l(h)}
\end{aligned}$$

$l(h) = -hE[f_{g(r)}] + ln(\frac{b_{g(r)} - E[f_{g(r)}]}{b_{g(r)} - a_{g(r)}}e^{ha_{g(r)}} + \frac{E[f_{g(r)}] - a_{g(r)}}{b_{g(r)} - a_{g(r)}}e^{hb_{g(r)}})$. Again rewriting the function $l(h)$ in terms of $l(h_{g(r)})$ where $h_{g(r)} = h(b_{g(r)} - a_{g(r)})$ and $s_{g(r)} = \frac{E[f_{g(r)}] - a_{g(r)}}{b_{g(r)} - a_{g(r)}}$. In this case $l(0) = 0$, $l'(0) = 0$ and $l''(0) \leq \frac{1}{4}$. Thus by Taylors theorem we have,

$$l(h) = l(h_{g(r)}) \leq \frac{1}{8}h^2(b_{g(r)} - a_{g(r)})^2$$

Hence from the above two equations and since e^z (where $z \in (-\infty, \infty)$) is a monotonic function we have,

$$E[Q_{g(r)}] \leq e^{\frac{1}{8}h^2(b_{g(r)} - a_{g(r)})^2} \quad (4)$$

Thus by equations 1, 2, 3 and 4 we have,

$$P[\bar{f} - E[\bar{f}] \geq t] \leq e^{-hNt + \frac{1}{8}h^2\sum_{i=1}^{N}(b_i - a_i)^2 + \sum_{i=1}^{k}(m_i - 1)\delta d_i h} \quad (5)$$

Minimizing the above convex function w.r.t. h we have,

$$h = \frac{4}{\sum_{i=1}^{N}(b_i - a_i)^2}(Nt - \sum_{i=1}^{k}(m_i - 1)\delta d_i)$$

but $h > 0$ and hence $t > \frac{\sum_{i=1}^{k}(m_i - 1)\delta d_i}{N}$. Substituting this value of h into equation 5 we prove the theorem,

$$P[\bar{f} - E[\bar{f}] \geq t] \leq e^{\frac{-2(Nt - \sum_{j=1}^{k}(m_j - 1)\delta d_j)^2}{\sum_{j=1}^{N}(b_j - a_j)^2}}$$

□

6. EXPERIMENTS

In this section we compare our bound to other competitive bounds that could be readily applied to the relational setting. In particular, we compare our STB with the i) Independent Test Bound (ITB) [9] and ii) Chromatic Test Bound (CTB) [15][3]. We first test how the STB behaves as a function of the strength of dependence (d) when the number of independent subsets (k) is small. The setting where k is small is more realistic and interesting since, large k implies that we are close to the iid setting in which case all the bounds have similar widths. We then test the bound widths on 2 real relational datasets.

6.1 Studying Trends

In these experiments we observe how the three bounds namely, CTB, ITB and STB behave with varying d.

We set k to be small which is the case with most real life relational datasets with the behavior of the three bounds for this setting being depicted in Figure 3. As we can see the STB is as tight as the i.i.d. bound when $d = 0$ and increases linearly with increasing d. However, the STB is tighter than the other two bounds for the most part except at very high levels of auto-correlation ($d \geq 0.9$). Note that for the CTB we considered the size of the largest independent subset to

[3]We do not apply the Chromatic PAC Bayes bound [26] since it is very difficult to apply it in practice. In particular, to apply it we have to first choose a relational classification algorithm, build the appropriate posterior (Q) on the hypothesis class represented by the algorithm and then choose the appropriate prior (P) on this class. It is not at all clear what this posterior or prior should be for the state-of-the-art relational classification algorithms. Moreover, the tightness of the bound would change for the same dataset depending on the algorithm and hence we would not be able to evaluate the quality of the bound just in terms of the properties of the dataset, which is the case for the other three bounds.

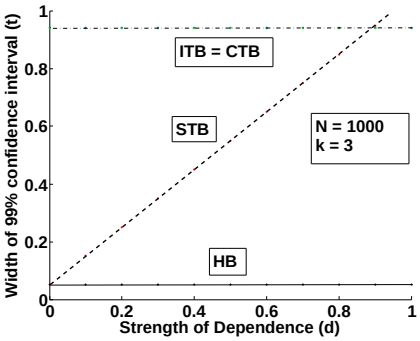

Figure 3: Comparison of bounds at small k with varying d. HB is the Hoeffding bound if the data were i.i.d. As we can see the STB outperforms other bounds everywhere except at very high d. Note that all values are rounded to two decimal places.

be 334 given that $k = 3$ and $N = 1000$. This gives the tightest possible bound for the given values of k and N.

6.2 Real Data Experiments

We now observe the behavior of the three bounds mentioned before on 2 real world datasets namely, Internet Movie Database (IMDB) (www.imdb.com) and WebKB [7]. Both of these datasets have been downloaded from the Alchemy website [16]. Note that if the width of a bound is greater than or equal to 1 we show the width to be 1 since the bound is in any case trivial.

IMDB: As the name suggests this dataset has information about movies, actors, directors etc. Given that we are evaluating test set bounds we choose only about 40% of the dataset on which to apply these bounds. The remaining 60% would generally be used for training some classification algorithm. With this, the test set size N turns out to be 110. *The classification task is to identify the gender of an actor based on the directors they have worked under.* Directors usually produce movies of a particular genre which may demand more actors of a certain gender. For example, action movies may have more male actors. The number of independent subsets k in this test set turns out to be 4 with the size of the largest independent subset being 55. The sizes of the 4 subsets are 55, 26, 14 and 15 with each having 39 males, 13 males, 12 males and 8 males respectively. The strength of dependence d estimated from the sample for this dataset is 0.1355.

As we can see in Figure 4, STB is much tighter than CTB or ITB. The reason for this tightness is due to the dependence of the STB on d and not just k. Hence, though k is small, a low value of d makes the STB more useful than the other two bounds.

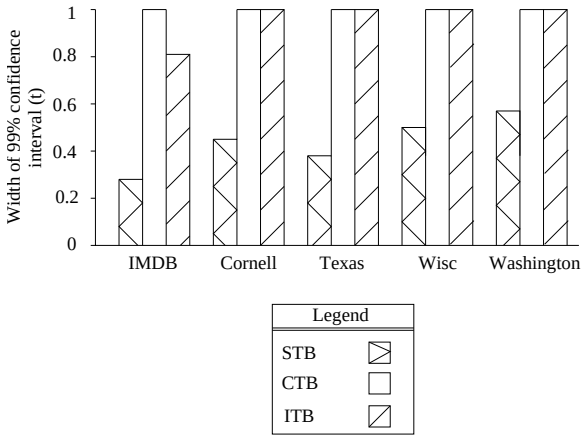

Figure 4: Comparison of bounds on two real world datasets namely, IMDB and WebKB. Cornell, Texas, Wisc (i.e. Wisconsin) and Washington are 4 datasets which together form the WebKB dataset.

WebKB: This dataset contains webpage and hyperlink information of 4 computer science departments. These are computer science departments in Cornell University, University of Texas, University of Wisconsin and University of Washington. The dataset has 4168 webpages which are categorized into 7 categories namely, student pages, faculty pages, departmental pages, instructor pages, course pages, members of project pages and research project pages. *The classification task is to identify student and non-student pages.* Usually when using this dataset people train on three of the departments webpages and test on the fourth. Hence, we have four plots for this dataset where each plot considers the corresponding department webpages as the test set. The size of each of these test sets is: $N = 867$ for Cornell (128 student pages), $N = 828$ for Texas (147 student pages), $N = 1267$ for Wisconsin (156 student pages) and $N = 1206$ for Washington (126 student pages). The estimated value for the strength of dependence for each of these test sets turns out to be: $d = 0.3961$ for Cornell, $d = 0.325$ for Texas, $d = 0.4617$ for Wisconsin and $d = 0.517$ for Washington. The number of independent subsets k is 1 for all of these test sets.

As we can see in Figure 4, STB is much tighter than CTB or ITB. In fact, both CTB and ITB are trivial (i.e. ≥ 1). Here again the low k and moderate d make the STB a more desirable alternative.

7. DISCUSSION

In the previous sections we have stated and derived a bound that depends on the strength of dependence between related datapoints. As we have seen the Hoeffding inequality is a special case of our bound when the datapoints are independent. The situations where our bound is particularly useful over the other existing bounds for relational data is when we have very few independent subsets of datapoints (i.e. low k) and the strength of dependence is low to moderate (i.e. d is close to 0). In these situations our bound would be able to exploit the weak dependency between these datapoints making it tighter than the existing bounds which (directly/indirectly) depend on k but not on d. Besides the two real datasets we applied our bound to, another strik-

ing real life example is social networking websites. Data from a social networking website is generally in the form of a huge graph with very few disjoint components (low k). Most of the people in such a network are linked through single and multiple hops to many other people very few of whom are close acquaintances (low d). It is also important to note that the central result in the paper, allows for multiple auto-correlation values for the same dataset. This means that recent research where auto-correlation is shown to be a local phenomenon [1] can be modeled in our framework.

It would be interesting in the future to derive bounds that are tighter than the ones derived here for high levels of dependence and low values of k but which reduce to known inequalities in the i.i.d. case. We believe however, that we have made a reasonable start in this endeavour.

8. REFERENCES

[1] P. Angin and J. Neville. A shrinkage approach for modeling non-stationary relational autocorrelation. In *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 707–712, Washington, DC, USA, 2008. IEEE Computer Society.

[2] M. Arias, A. Feigelson, R. Khardon, and R. Servedio. Polynomial certificates for propositional classes. *Inf. Comput.*, 204(5):816–834, 2006.

[3] G. Bennett. Probability inequalities for the sums of independent random variables. *JASA*, 57:33–45, 1962.

[4] A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Computational Learing Theory*, pages 203–208, 1999.

[5] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[6] W. Cohen. Polynomial learnability and inductive logic programming: Methods and results. *New Generation Computing*, 13:369–409, 1995.

[7] M. Craven and S. Slattery. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning*, 43(1-2):97–119, 2001.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[9] A. Dhurandhar and A. Dobra. Distribution free bounds for relational classification. *Knowledge and Information Systems*, 31:55–78, 2012.

[10] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1309, 1999.

[11] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[12] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford, 3 edition, 2001.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *JASA*, 58(301):13–30, 1963.

[14] G. Hulten, P. Domingos, and Y. Abe. Mining massive relational databases, 2003.

[15] S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24:234–248, 2004.

[16] S. Kok, P. Singla, M. Richardson, and P. Domingos. The alchemy system for statistical relational ai. Technical report, Department of Computer Science and Engineering, UW, http://www.cs.washington.edu/ai/alchemy/, 2005.

[17] L. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36:2126–2158, 2006.

[18] J. Langford. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, 6:273–306, 2005.

[19] D. McAllester. Pac-bayesian model averaging. In *In Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170. ACM Press, 1999.

[20] D. McAllester. *Generalization Bounds and Consistency, chapter in book Predicting Structured Data*. The MIT Press, 2007.

[21] M. Mohri and A. Rostamizadeh. Stability bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems 20*, pages 1025–1032. MIT Press, Cambridge, MA, 2008.

[22] J. Neville. Statistical models and analysis techniques for learning in relational data. Ph.D. Thesis, University of Massachusetts Amherst, 2006.

[23] M. Okamoto. Some inequalities relating to the partial sum of binomial probabilites. *Annals of the institute of Statistical Mathematics*, 10:29–35, 1958.

[24] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 3 edition, 1991.

[25] L. Raedt. First order jk-clausal theories are pac-learnable. *Artificial Intelligence*, 70:375–392, 1994.

[26] L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic pac bayes bounds for non-iid data. In *Twelfth International Conference on Artificial Intelligence and Statistics*. Omnipress, 2009.

[27] M. Richardson and P. Domingos. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136, 2006.

[28] J. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-hoeffding bounds for applications with limited independence. *SIAM J. Discrete Math*, 8:223–250, 1995.

[29] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *In Proc. 18th Conference on Uncertainty in AI*, pages 485–492, 2002.

[30] V. Vapnik. *Statistical Learning Theory*. Wiley & Sons, 1998.