Test Set Bounds for Relational Data that vary with Strength of Dependence

AMIT DHURANDHAR
IBM T.J. Watson
and
ALIN DOBRA
University of Florida

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 on the generalization error of a classifier 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. Moreover, the bound provides insight in the computation of effective sample size which is an important notion introduced by [Jensen and Neville 2002].

Categories and Subject Descriptors: H.2.8 [**Data Management**]: Data Mining

General Terms:

Additional Key Words and Phrases: Statistical Relational Learning, distribution free bounds

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 subarea of Machine Learning namely, Statistical Relational Learning (SRL) [Getoor and Taskar 2007] is concerned with modeling of uncertainty in such type of non-i.i.d. or relational data.

A. Dhurandhar, IBM T.J. Watson, Yorktown Heights, NY-10598, USA.
A. Dobra, University of Florida, Gainesville, Fl-32611, USA.

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 [Richardson and Domingos 2006; Taskar et al. 2002; Friedman et al. 1999; Neville 2006] 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. [Hulten et al. 2003; Getoor and Taskar 2007] expressed a need for developing such a theory.

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 [Markov 1890; Papoulis 1991; Grimmett and Stirzaker 2001], the Chebyshev inequality [Chebyshev 1859; Papoulis 1991; Grimmett and Stirzaker 2001] and the Hoeffding inequality [Hoeffding 1963] 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 [Hoeffding 1963]. Other such inequalities are given by Chernoff [Chernoff 1952], Bennett [Bennett 1962] and Okamoto [Okamoto 1958]. Distribution-free bounds on the generalization error of a classifier are provided by [Vapnik 1998] based on a property of the classifier space called Vapnik-Chervonenkis (VC) dimension. [Devroye et al. 1996] provided distribution-free bounds for the k-nearest neighbor algorithm. [McAllester 2007] improved the Probably Approximately Correct (PAC) Bayes bounds for linear decoders. These bounds are tighter than the ones previously introduced [McAllester 1999]. [Blum et al. 1999] provided bounds 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 by [Langford 2005]. 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. The reason they are easier to apply is that, they depend on the test error and the number of test samples and are oblivious to the specifics of the classification algorithm (such as properties of the hypothesis space it spans) used to make predictions. In this paper, we derive 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) [Kontorovich and Ramanan 2006; Mohri and Rostamizadeh 2008]. 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 ($k \leq n$). [Schmidt et al. 1995] provides bounds assuming k-wise independence, 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 [Cohen 1995; Raedt 1994; Arias and Khardon 2002; Arias et al. 2006]. The learnability results are primarily based on two formal models of learning namely, PAC learning and learning from equivalence and membership queries.

Bounds for applications we are interested in have been derived by [Dhurandhar and Dobra 2011; Janson 2004; Ralaivola et al. 2009]. However, 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 (or weakly) correlated the interacting datapoints are. In this paper, we derive bounds that vary with the degree of dependence between related datapoints, which is an attractive and potentially very useful feature. The situations where our bound would be particularly useful over these other bounds is when most of the datapoints are correlated but this correlation is weak. In these situations our bound would be able to exploit the weak dependency between these datapoints making it tighter than the existing bounds. One striking real life example of this situation is social networking websites. Data from a social networking website is generally in the form of a huge graph with very few disjoint components (i.e. most of the data is correlated). 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 dependence). Hence, even if all the datapoints are (seemingly) dependent our bound will be as tight as an i.i.d. bound when this dependence is weak. Moreover, as we will see later our bound becomes the standard Hoeffding or McDiarmids inequality when the datapoints are independent.

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 formally define strength of dependence and its relation to relational auto-correlation. In Section 4, we clearly state and justify the assumptions needed in obtaining the results.

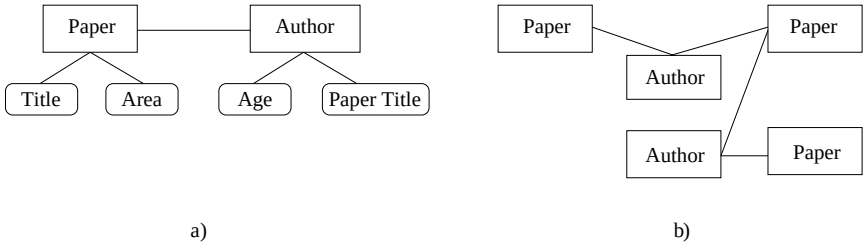

Fig. 1. a) represents a relational schema with object types, *Paper* and *Author*. The relationship between them is many-to-many. The rounded boxes linked to these object types denote their respective attributes. b) is the corresponding data graph which shows authors linked to the papers that they authored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

In Section 5 we state these results. The proof of the main result is given in the appendix. In Section 6, we compare our bound with other relevant bounds in literature in terms of ease of use and conditions under which each of them is tight i.e have a value that is close to the value of the bound if the data were i.i.d. In Section 7, we test the validity of our central assumption and observe the sensitivity of the bound to violation of this assumption. We then evaluate the bounds on some real life relational datasets. In Section 8, we use information from the bound to derive upper and lower bounds on the effective sample size. We summarize the findings in this paper and suggest future lines of research in section 9.

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 5 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

Fig. 2. Relational database representation of the relational dataset in Figure 1b. The table on the upper left contains objects of type *Paper* and the table on the right contains objects of type *Author*. The attribute Title is a primary key in *Paper* and the attribute Paper Title is the corresponding foreign key in *Author*. The table at the bottom is the Universal table formed by joining these 2 tables.

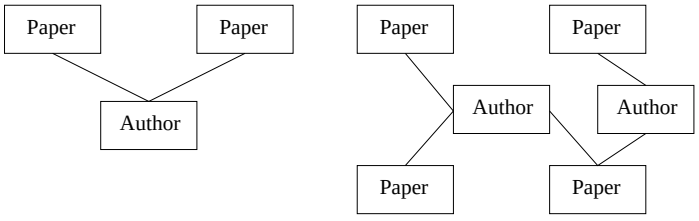

Fig. 3. Above we see a disconnected data graph with 2 components of variable size (i.e. size 2 component on the left and size 4 component on the right).

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) [Neville 2006], 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.

Another popular representation of relational data is in Relational Database Management Systems (RDBMS) where object and link information is stored in tables. A single table stores objects or link information of a particular *type*. The columns of such a table denote the attributes associated with the particular *type* while each row stores information of each individual object (or link) of that *type*. This is shown in Figure 2, which represents a relational dataset that identical to the one represented by the data graph 1b. Though the data is stored in separate tables to avoid redundancy, the information in these tables can be put into a single large table called the Universal table [Malvestuto 1989]. Thus, the *Universal table has all the information that is present in the relational dataset*. This view of a relational dataset is critical in understanding the generality of the joint distributions we consider in this paper.

The procedure of combining smaller tables to form larger tables is called a Database Join (or just a Join)[1]. Usually a Join is between the primary key in one table and the foreign key in another table. A primary key is an attribute/set of attributes that uniquely identify a row in their table (e.g. Title in *Paper*) and a foreign key is an attribute/set of attributes that uniquely identify a row in a different table (e.g. Paper Title in *Author* uniquely identifies rows in *Paper*).

Probabilistic Models over Relational Data: Probabilistic Models over relational data (PMRD) [Getoor and Taskar 2007] are structured graphical models that are used to handle uncertainity in relational domains. 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

[1]Joins are of different types but their discussion is unnecessary for this paper.

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 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 3. The joint distribution over the data graph in Figure 3 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 $P[A_3, A_4, A_5, A_6]$.

Another solution for handling multiple values for a single object without aggregating them, is creating multiple copies for that object. This is seen in Figure 2, where we have a Universal table view of the corresponding relational dataset. We see here, that "paper2" has two entries in the table. The corresponding joint distribution for this type of modeling would be,

$$P[B_1, B_2, B_3, B_4] = P[B_1, B_2]P[B_3, B_4]$$

where B_i denotes the attribute set $\{Title^i, Area^i, Age^i\}$ for the i^{th} row in the Universal table 2.

In either case, 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.

It also important to 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. It maybe the case sometimes that the distribution of pairs of sets of attributes is the same i.e. $P[A_1, A_2] = P[A_3, A_4] = ... = P[A_{N-1}, A_N]$ or combinations of larger sets, since those are the basic templates, however, all these cases imply $P[A_i] = P[A_j]$ $\forall i, j \in \{1, ..., N\}$. Realize that this 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.

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 that takes as input two parameters and outputs values in the range $[0, M]$ where M is a positive integer. Such a function generally outputs a 0 if the two parameters are equal and some positive real value if they are unequal. Common examples of loss functions used in classification are 0-1 loss, hinge loss, least squares loss, etc. The expected value of $\lambda(\zeta(x), y)$ over $X \times Y$ space is defined as the generalization error of the particular classifier ζ. Formally,

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

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.

3. STRENGTH OF DEPENDENCE

Strength of dependence d, measures the degree of statistical dependence of an attribute on related/linked datapoints. In the case of relational data, this statistical dependence is called relational auto-correlation and it measures the degree of similarity between values of the same attribute on related datapoints. As is the case with computing cross-correlation in i.i.d. domains there is no standard metric for computing auto-correlation in relational domains either. We discuss here some commonly used metrics for computing relational auto-correlation.

As mentioned by [Neville 2006], relational auto-correlation (ρ) for a continuous attribute can be estimated as follows:[2],

$$\rho = \frac{\sum_{\{i,j\} \in R} (z_i - \bar{Z})(z_j - \bar{Z})}{\sum_{i \in V_R} n_i (z_i - \bar{Z})^2}$$

where N is the size of the dataset, $\forall i \in \{1, ..., N\}$ $z_i \in Z$ are the values of the continuous attribute, $\bar{Z} = \frac{1}{N} \sum_{i=1}^{N} z_i$, R is a set of all pairs of indices of the datapoints that are related in the dataset, V_R is the set of indices of the datapoints that occur in some element of R and n_i is half of the number elements of R in which i occurs. The n_i are required for normalization allowing the value of ρ to lie only in the interval $[-1, 1]$.

[2]For other applications ρ may be estimated differently, but the results in this paper will still apply if the problem can be expressed in our setting.

If the attribute is discrete there are a number of normalized metrics such as ϕ coefficient, Pearsons contingency coefficient, Cramers V [Sachs 1984] etc. However, all of these metrics can be used only if the attribute is binary. In the general case where the attribute may have higher cardinality a normalized version of the Kullback Leibler divergence [Kullback and Leibler 1951] (or relative entropy or information gain as it is sometimes called) can be used. Thus, ρ for a discrete attribute can be estimated as follows,

$$\rho = \frac{\sum_{i=1}^{k} KL(p_i||q)}{kH_q}$$

where k is the number of independent subsets, p_i is the empirical distribution over the values of the discrete attribute computed on the i^{th} independent subset, q is the maximum entropy distribution over the values of the discrete attribute, KL(.) is the Kullback Leibler divergence and H_q is the entropy of q. To get a better feel for what each of the terms in formula mean, we give the following example. Consider $k = 2$ and the attribute Z takes 3 values z_1, z_2, z_3. Let the first independent subset have 12 datapoints with 5 taking the value z_1, another 5 taking the value z_2 and the remaining 2 taking the value z_3. In this case $p_1 = (\frac{5}{12}, \frac{5}{12}, \frac{2}{12})$. Let the second independent subset have 15 datapoints with 3 taking the value z_1, another 7 taking the value z_2 and the remaining 5 taking the value z_3. In this case $p_2 = (\frac{3}{15}, \frac{7}{15}, \frac{5}{15})$. The distribution with the maximum entropy over the 3 values of Z is given by, $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Consequently, ρ in this case is, 0.114 indicating that the auto-correlation in this dataset is low.

In the result that we present in Section 5, we are not concerned with the sign of the auto-correlation (i.e. positive or negative), but rather its numerical value. In other words, we are only concerned about how strong (or weak) the dependence is or how far (or close) we are from independence. Consequently, we define strength of dependence as follows,

$$d = |\rho|$$

where $|.|$ stands for absolute value. The value of d lies in the interval $[0, 1]$ where 0 means the datapoints are not correlated while 1 means that the datapoints are highly correlated.

4. 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[3]. 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.[4]

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

4.1 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,

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

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.

At $d = 1$ the conditional expectation is equal to the average value of the variables it is conditioned on, which essentially are all the same (since d is 1) and hence the conditional expectation in this scenario can be written as, $E[Z_i|Z_{i-1} = z_{i-1}, ..., Z_1 = z_1] = z_{i-1}$ $\forall i \in \{2, ..., m\}$. This means that at $d = 1$ the sequence $Z_1, ..., Z_m$ can be viewed as a martingale.

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.

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

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

4.2 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 [Ralaivola et al. 2009; Dhurandhar and Dobra 2011]. 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.

5. 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. The proof of Theorem 1 is in the appendix.

The result in Theorem 1 is for the general setting where random variables (with possibly different ranges) are partitioned into statistically independent sets such that random variables in a particular set interact with some/all others in the same set and each of these independent sets has its own strength of dependence parameter.

We now introduce some of the notation that is used in the statement of the theorem. 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 auto-correlation between datapoints in each subset (d_j where $j \in \{1, ..., k\}$) and the ranges of Z_i ($[a_i, b_i]$). The bound is tight when N is large and k is close to N or when N is large and the d_j $j \in \{1, ..., k\}$ are low. 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 non-trivial (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 $d_j = 0$ $\forall j \in \{1, ..., k\}$ or if $k = N$ 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 classification we usually have a single strength of dependence parameter d (or auto-correlation parameter) 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) – where x_i corresponds to the input consisting of relevant input attributes and class labels (viz. that of neighbors) while $y_i \in Y$ – 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 $\lambda_i = \lambda(\zeta(x_i), y_i)$ $\forall i \in \{1, ..., N\}$.

Bound	Difficulty in applying	Becomes tighter when
CPB	High	Large N **and** low χ^* (or χ^*_u) **and** appropriate P
CTB	Low	Large N **and** low χ^* (or χ^*_u)
ITB	Low	Large k (number of independent components)
STB	Low	Large k **or** low d

Table I. Comparison of inequalities in terms of ease of use and conditions under which they are tight. χ^*_u denotes the upper bounds on χ^*. Generally when applying CPB or CTB one uses some χ^*_u since the computation of χ^* is NP-hard.

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. As we can see, the above bound is simple to apply to a relational dataset since the required parameters (d, N, M and k) are not too difficult to obtain.

6. COMPARISON OF BOUNDS

In this section we describe and evaluate bounds derived by [Dhurandhar and Dobra 2011; Janson 2004; Ralaivola et al. 2009] along with the bound presented in this paper. These three bounds are applicable in the settings that our bound is applicable and hence it is important to discuss their trade-offs. Table I provides a comparison of these bounds in terms of applicability and tightness.

6.1 Chromatic PAC Bayes bound (CPB)

The Chromatic PAC Bayes bound [Ralaivola et al. 2009] is a generalization of the standard PAC Bayes bound [McAllester 2007] for non-i.i.d. data. Like the PAC Bayes bound the CPB is a training-set bound which bounds the training error to the generalization error of a stochastic classifier (called Gibbs classifier). Given a posterior distribution Q over a hypothesis space the Gibbs classifier classifies an input by randomly picking a hypothesis from this space according to Q and then outputting the prediction of this hypothesis on the input. We now state the CPB theorem,

THEOREM 3. *[Ralaivola et al. 2009]* $\forall N$, $\forall D_N$, $\forall H$, $\forall \delta \in (0, 1]$, $\forall P$, *with probability at least* $1 - \delta$ *over the random draw of* $(X \times Y)^N \sim D_N$, *the following holds,*

$$\forall Q, kl(\hat{e_Q}||e_Q) \leq \frac{\chi^*}{N}[KL(Q||P) + \ln \frac{N + \chi*}{\delta \chi^*}]$$

where N is the sample size, D_N is a distribution over all samples of size N, H denotes the hypothesis space, P is a prior over this hypothesis space, Q is a poste-

rior over this hypothesis space, kl(.) is the symmetric Kullback Leibler divergence, KL(.) is the asymmetric Kullback Leibler divergence, χ^* is the fractional chromatic number, $\hat{e_Q} = E_{h\sim Q}\frac{1}{N}\sum_{i=1}^{N} I[h(x_i) \neq y_i]$ and $e_Q = E_{(X\times Y)^N \sim D_N}\hat{e_Q}$.

In the above bound, the fractional chromatic number χ^* is to be computed for a graph called the dependency graph whose vertices are the sample points and there is an edge between two vertices if and only if the two points are dependent. This graph can be formed from relational data by observing the dependencies in the data graph. However, χ^* is generally very difficult to find since computing it for arbitrary graphs is NP-hard. The chromatic number χ which is the minimum number of colors needed to color the vertices of a graph such that adjacent nodes have different color is an upper bound on χ^* and hence can replace χ^* in the above bound but it too is (NP-) hard to compute. However, the fact that any upper bound on χ^* can be used to replace χ^* is extremely useful. As such we have the following set of inequalities,

$$1 \leq \phi \leq \chi^* \leq \chi \leq \Delta + 1 \tag{1}$$

where ϕ is the size of the largest clique in the graph and Δ is the degree of the maximum degree vertex in the graph.

Applicability: CPB is generally very difficult to apply to arbitrary classification algorithms/classes of functions in practice. The reason for this is that building the appropriate Q requires a great deal of sophistication [Langford 2005]. Moreover, from bayesian approaches we know how difficult it is to choose the appropriate prior P. In addition to all of this, the bound we get applies to the Gibbs classifier which is likely to be different from the classifier we are interested in and hence we need to relate the GE of these two classifiers so as to get relevant results. Hence, it is reasonable to assume that this bound will be used by practitioners only for those function classes where all this work has already been done.

Tightness: The bound is tight when P is chosen carefully to be close to Q (in terms of KL(.)), N is large and χ^* or its upper bounds (that we will actually use) are much less than N.

6.2 Chromatic Test Set Bound (CTB)

We refer to the bound presented by [Janson 2004], relevant to our setting as the Chromatic Test Set Bound. CTB is a Hoeffding style bound extended to the non-i.i.d. setting. Similar to CPB this bound also depends on the fractional chromatic number χ^* of a graph. However, unlike the CPB, the CTB is a test-set bound and does not depend on the details of the algorithm it is applied to. The CTB for the relational setting can be written as follows, for $t > 0$

$$P[|HE - GE| \geq t] \leq 2e^{\frac{-2Nt^2}{\chi^* M^2}}$$

where M is the range of the loss function.

Applicability: Here again as in CPB, χ^* is difficult to compute and hence upper bounds on χ^* that are easy to compute as shown in equation 1 may be used. In fact, in our setting given a data generation model of the form $P[Z_1, ..., Z_N] = \prod_{i=1}^{k} T_i$ where $k \in \{1, ..., N\}$ is the number of disjoint independent subsets of the random variables each of size m_i, we can (tightly) upper bound χ (and hence χ^*) by the size of the largest independent subset. Formally, $\chi \leq max_{i \in \{1,...,k\}} m_i$. With this, CTB is easy to apply. Note that $\phi = \chi^* = \chi = max_{i \in \{1,...,k\}} m_i$ when all the random variables in the largest independent subset are pairwise correlated.

Tightness: The bound is tight when N is large and χ^* or the upper bound we use for it is small.

6.3 Independent Test Set Bound (ITB)

We call the bound presented by [Dhurandhar and Dobra 2011] an Independent Test Set Bound since the bound depends on the number of independent subsets of the datapoints k in the dataset. Given a data generation model of the form $P[Z_1, ..., Z_N] = \prod_{i=1}^{k} T_i$ where $k \in \{1, ..., N\}$ is the number of disjoint independent subsets of the random variables each of size m_i and l is the least common multiple of $m_1, ..., m_k$, we have for $t > 0$

$$P[|HE' - GE| \geq t] \leq 2e^{\frac{-2kt^2}{M^2}}$$

where M is the range of the loss function $\lambda(.,.)$, $HE' = \frac{1}{lk} \sum_{j=1}^{k} \frac{l}{m_j} e_j$, $e_j = \sum_{r=f_j}^{f_j+m_j} \lambda_r$, f_j is the index of the first variable in T_j and λ_r is the loss of the classifier for the r^{th} datapoint.

This inequality bounds not the hold-out error but rather a weighted version of the hold-out error where the weights depend on the size of the independent subsets. When all the subsets are of the same size this error becomes the hold-out error. In either case this error can be easily computed from the test set.

Applicability: This bound is easy to apply in practice.

Tightness: The bound is tight when k is large. This bound has an interesting relation to CTB. As mentioned before, for the CTB χ^* or χ are difficult to compute and hence we can use $L = max_{i \in \{1,...,k\}} m_i$ which is an upper bound on both of these quantities in the CTB. As it turns out $\frac{N}{L} \leq k$ and hence ITB will always be at least as tight as CTB. The equality holds when all the m_i $i \in \{1, ..., k\}$ are the same.

6.4 Strength of Dependence Test Bound (STB)

This is the bound presented in this paper. The name follows from the fact that among other things the bound depends on strength of dependence.

Applicability: This bound is easy to apply in practice.

Tightness: The bound is tight when k is large or when d is small. Thus, unlike the

other bounds this bound can be tight if either of the conditions is satisfied making it more useful in real life.

7. EXPERIMENTS

The goal of this section is to get a more "hands-on" feel for the STB (our bound) in terms of how it behaves compared to the other bounds that we have mentioned. In addition, we also try to justify assumption 1 through synthetic data experiments and a robustness (or sensitivity) study in an attempt to validate the use of the derived bound.

We divide this section into the following 4 parts.

—In the first part, we observe how the STB behaves as a function of the strength of dependence (d) when the number of independent subsets (k) is small and then when the number of independent subsets is large. The setting where k is small is the most challenging in terms of tightness for the other bounds in literature and hence we want to compare the performance of our bound particularly in this setting.

—In the second part, we study how erroneous assumption 1 is likely to be, by reporting the mean squared error between the true conditional expectations and the value assigned to them as a result of the assumption, for different values of d.

—In the third part, we perform a robustness analysis to observe the sensitivity of the bound to assumption 1. We do this by reporting the ratio of the width of the bound with error to the width of the bound without error and plot this value as a function of the error.

—In the fourth part, we apply our bound along with some of the other bounds mentioned to real life relational datasets and observe how tight the results are. The bounds that we apply in the first and the fourth part are CTB, ITB and STB. We do not apply CPB since as mentioned in the previous section 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 such as Markov Logic Networks [Richardson and Domingos 2006] or Relational Dependency Networks [Neville 2006] etc. 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 3 bounds.

7.1 Studying Trends

In these experiments we observe how the three bounds namely, CTB, ITB and STB behave with varying d. First, 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 4. 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

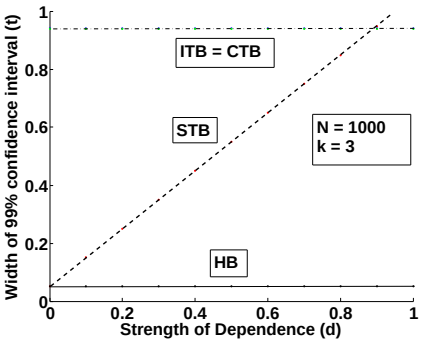

Fig. 4. Comparison of bounds at small k with varying d. HB is the Hoeffding bound if the data were i.i.d. As we can see the STB outperforms other bounds everywhere except at very high d. Note that all values are rounded to two decimal places.

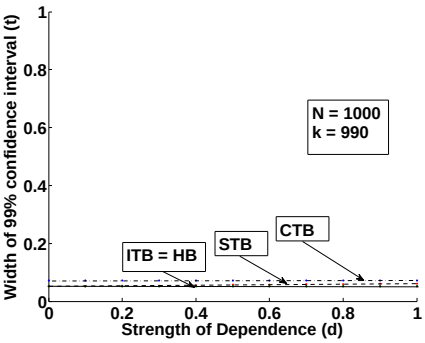

Fig. 5. Comparison of bounds at large k with varying d. HB is the Hoeffding bound if the data were i.i.d. In this case all the bounds are more or less equally tight. Note that all values are rounded to two decimal places.

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

Second, we set k to be large as depicted in figure 5. Here, we observe that all the bounds are very close to the i.i.d. bound. Hence, using anyone of the three bounds seems equally good. Note for the CTB we set the size of the largest independent subset to be 2 which again gives the best bound for the particular k and N.

7.2 Violation of Assumption 1

In this subsection, we observe the extent to which assumption 1 may be violated on average. The synthetic distributions we consider, generate data where there is significant variability in local auto-correlation i.e. certain components have zero auto-correlation while some others have a auto-correlation of 1. We do this to test the degree of violation of assumption 1 under such extreme circumstances. As in the previous study, the synthetic distributions span the space of low k and high k for different values of d. We set the number of classes to 2 and the number of interacting datapoints/variables to 100 (i.e. $N = 100$) for all the distributions. Hence, the maximum entropy distribution (q) over the labels is $(0.5, 0.5)$. The number of explanatory attributes is set to 5, where each attribute takes 2 values.

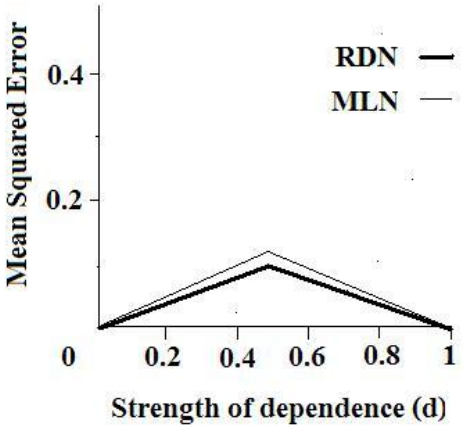

Fig. 6. Mean squared error between the true conditional expectation and the value computed from assumption 1 with varying d.

As mentioned before, since a relational dataset can be represented as a Universal table where the columns are attributes, we can consider these 5 attributes to be the once chosen from some relational dataset, with them belonging to the same or different object types.

Joint Distributions: For the first distribution, we set $k = 10$ (i.e. low k) where 5 components are of size 5 and the other 5 components are of size 15. For two of the size 5 components and one of the size 15 component we set distribution over the labels to be $(1, 0)$ i.e. all assignments to the variables in the joint probability distribution, which indicate class 2 have 0 probability, while the remaining assignments have equal probability for these components. For the remaining 7 components the distribution over the labels is $(0.5, 0.5)$ i.e. in each of these 7 components, assignments to variables where half belong to a particular class have equal probability and sum to 1. This setup for the first distribution gives a strength of dependence value of $d = 0.3$. For the second distribution, we have the same number and sizes of components as before, but for three of the size 5 components and two of the size 15 components we set the distribution over the labels to be $(1, 0)$. For the remaining 5 components the distribution over the labels is $(0.5, 0.5)$. This setup for the second distribution gives a strength of dependence value of $d = 0.5$. For the third distribution, we again have the same number and sizes of components as before, but for two of the size 5 components and one of the size 15 components we set the distribution over the labels to be $(0.5, 0.5)$. For the remaining 7 components the distribution over the labels is $(1, 0)$. This setup for the third distribution gives a strength of dependence value of $d = 0.7$. We create three more distributions which have the preceding values for d, but which have high k. For next three distributions, we set $k = 50$ where each component is of size 2. With this, in the fourth

distribution for 15 of the components we set the distribution over the labels to be $(1, 0)$ and for the remaining 35 components the distribution is $(0.5, 0.5)$. This gives a value of $d = 0.3$. In the fifth distribution, for 25 of the components we set the distribution over the labels to be $(1, 0)$ and for the remaining 25 components the distribution is $(0.5, 0.5)$. This gives a value of $d = 0.5$. In the sixth distribution, for 35 of the components we set the distribution over the labels to be $(1, 0)$ and for the remaining 15 components the distribution is $(0.5, 0.5)$. This gives a value of $d = 0.7$.

Classification Models: We train and test using state-of-the-art collective classification models namely, a Markov Logic Networks (MLN) [Richardson and Domingos 2006; Domingos and Richardson 2004] and Relational Dependency Networks (RDN) [Neville 2006]. The MLN is learned discriminatively using the tool Alchemy [Kok et al. 2005]. We then perform Maximum aposteriori (MAP) inference to get predictions. For RDNs the inference is done using Gibbs sampling over the learned conditional probability distributions.

Training and Testing: The training set size is 10000 for each distribution and the test set size is 5000. For each distribution, we find the mean squared error between the conditional expectations and the value computed using assumption 1 (all the expectations are approximated by their corresponding sample averages). We then average these errors for distributions with the same d. Thus, the final mean squared error computed for a particular d takes into account low and high values of k. The behavior and amount of error for different values of d is seen in figure 6. From the figure we see that the assumption is most violated at $d = 0.5$ by a margin of about 10% (0.13 to be exact). The error does not seem to be large however, the gravity of the violation of assumption 1 on the tightness of the bound can only be judged by the sensitivity of the bound to this error. This is precisely what we do in the next section.

7.3 Sensitivity of STB to Assumption 1

In this subsection, we study the sensitivity of the bound to assumption 1. To accomplish this, we rewrite assumption 1 as follows:

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

where $\epsilon_i \geq 0 \; \forall i \in \{2, ..., m\}$ are the errors in the assumption 1. With this modified assumption and with essentially the same proof when using assumption 1 we can derive a new upper bound which is: $P[|HE - GE| \geq t] \leq 2e^{\frac{-2(Nt-(N-k)M(d+\epsilon))^2}{NM^2}}$ where $\epsilon = max_i(\epsilon_i)$.

We now define α sensitivity as follows:

$$\alpha \text{ sensitivity} = \frac{\text{width of } \alpha\% \text{ confidence bound with error}}{\text{width of } \alpha\% \text{ confidence bound without error}}$$

We set $N = 100$, $M = 1$ and compute $\alpha = 99\%$ sensitivity for each k sampled

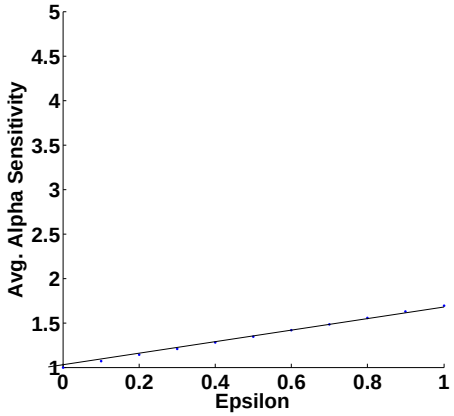

Fig. 7. The behavior of Average $\alpha = 99$ Sensitivity with increasing error (ϵ) is shown above.

(50 samples) uniformly from $\{1, 2, ..., 100\}$, d sampled uniformly from $[0, 1]$ (100 samples) for each $\epsilon \in \{0, 0.1, 0.2, ..., 1\}$. We then average the α sensitivity for each epsilon and report this value as a measure of sensitivity/robustness of the bound to assumption 1. This is seen in figure 7. We see that the sensitivity is a linear function of ϵ with a small slope. When $\epsilon = 1$, which is the highest error (and probably unlikely), the average α sensitivity is 1.7, which implies the width of the 99% confidence interval with error, on average, is less than twice as much as the width without error. This means that the derived bound is not too sensitive to the violation of assumption 1.

Remark: From subsection 7.2 we see that the worst ϵ is likely to be around 0.13 and combining this information with the sensitivity analysis we can infer that the bound will most likely be valid and useful in practice.

7.4 Real Data Experiments

We now observe the behavior of the three bounds on 2 real world datasets namely, Internet Movie Database (IMDB) (www.imdb.com) and WebKB [Craven and Slattery 2001]. Both of these datasets have been downloaded from the Alchemy website [Kok et al. 2005]. 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.

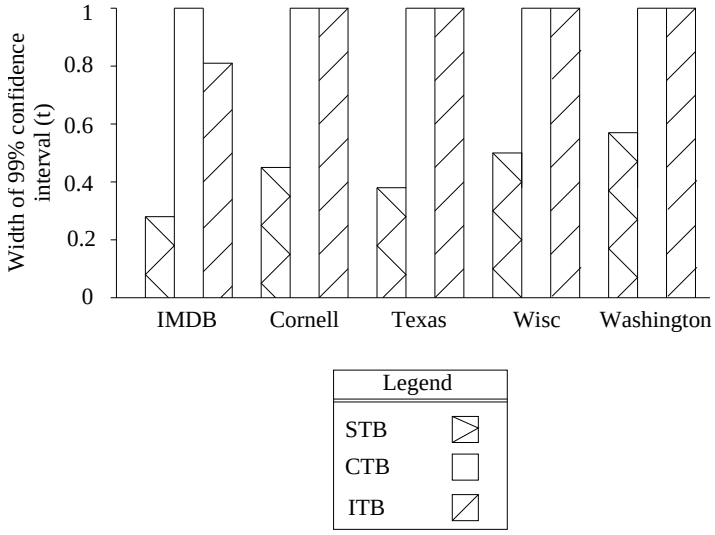

Fig. 8. Comparison of bounds on two real world datasets namely, IMDB and WebKB. Cornell, Texas, Wisc (i.e. Wisconsin) and Washington are 4 datasets which together form the WebKB dataset.

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 8, 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.

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 8, 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.

8. BOUNDS ON EFFECTIVE SAMPLE SIZE

The notion of effective sample size (γ) was introduced by [Jensen and Neville 2002]. Relational data exhibits auto-correlation and hence has less information than an i.i.d. sample of the same size. γ is essentially the size of an i.i.d. sample which has the same amount of information as the relational dataset at our disposal. Hence, if the relational dataset of size N has no dependencies then the γ would just be N. At the other end of the spectrum if the auto-correlation in the relational dataset of size N is 1, then γ would be 1 since, there is just one datapoint worth of information in the dataset. It is not clear however what the value of γ is between these extreme cases. [Jensen and Neville 2002] built an estimator for this quantity based on intuitions and simulated the value of this quantity with varying amounts of auto-correlation and connectivity in the graph (i.e. k). In this paper however, lower and upper bounds can be derived for γ as a direct consequence of the theory we have developed rather than plain intuition. This is another benefit of the results we have derived.

Lower Bound on γ: The lower bound on the γ is k – the number of independent subsets [Dhurandhar and Dobra 2011]. This is easy to see since if we have k independent sets of datapoints and if the correlation within each of these sets is maximum (i.e. 1), we essentially have a dataset with k independent datapoints. Thus,

$$\gamma \geq k$$

Upper Bound on γ: The upper bound is more interesting and less obvious to derive. To derive the upper bound we equate the bound in Theorem 2 with the corresponding i.i.d. bound based on γ. We thus have,

$$
\begin{aligned}
e^{\frac{-2\gamma^2 t^2}{NM^2}} &= e^{\frac{-2(Nt-(N-k)Md)^2}{NM^2}} \\
\gamma^2 t^2 &= (Nt-(N-k)Md)^2 \\
\gamma &= N - \frac{(N-k)Md}{t} \\
&\leq N-(N-k)d \\
&= k+(N-k)(1-d)
\end{aligned}
$$

where we used the fact that $|HE - GE| \leq M$ and hence the question of $P[|HE - GE| \geq t]$ is only sensible to ask for $0 < t \leq M$. Thus, restricting the value of t to the interval $(0, M]$ gives us our upper bound on γ.

With this we have the following set of inequalities,

$$k \leq \gamma \leq k + (N-k)(1-d)$$

It is interesting to note that as $k \to N$ both the upper and lower bounds converge to N for a fixed d. When $d \to 1$ the upper bound converges to the lower bound k. The gap between upper and lower bounds is large only when $d \to 0$ for a $k < N$. However, we know when $d = 0$, $\gamma = N$ i.e. the upper bound becomes the true value of γ.

In the future it would be interesting to derive a tighter lower bound for γ which depends also on d rather than just k.

9. CONCLUSION

In this paper we have presented a bound that varies with 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. We have compared our bound with other related bounds in terms of ease of use, tightness and performance on real datasets. As it turns out our bound does well with respect to these metrics. In addition to being useful, our bound also provides insight in the estimation of effective sample size.

It is important to note that the central result in the paper, allows for multiple auto-correlation values for the same dataset. This means that the current research where auto-correlation is shown to be a local phenomenon [Angin and Neville 2008] can be modeled in our framework. We know that our bound tightens with increasing k for a given d. One possible way of choosing a more ambitious k could be by limiting the length of the connections, as is generally done when building relational models. Another way could be by thresholding the auto-correlation and splitting regions with low local auto-correlation. However, as we have seen, one of the major strengths of the bound is that it tightens with high k *or* low d, which implies that there are two knobs to tighten the bound as opposed to just one (k). This means that even a coarse estimate of k such as number of disconnected components can provide a reasonable bound as it might lead to a low d.

It would be interesting in the future to derive bounds that are tighter than the ones derived here for high levels of dependence and with few independent subsets but which reduce to known inequalities in the i.i.d. case. It would also be interesting

to include other forms of dependence (i.e. dependence between attributes besides the target) than just auto-correlation, when deriving such bounds. Incorporating these other dependencies would in all likelihood lead to better bounds than the ones we have here, in the same way as incorporating auto-correlation has led to improvement over previously existing bounds. We believe, however, that we have taken an initial but firm step forward in deriving bounds that take into account such dependencies and which, as we have seen can be very useful in practice.

REFERENCES

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

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

Arias, M. and Khardon, R. 2002. Learning closed horn expressions. *Inf. Comput. 178,* 1, 214–240.

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

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

Chebyshev, P. 1859. Sur les questions de minima qui se rattachent la reprsentation approximative des fonctions. In *Mm. Acad. Sci. Ptersb. 7*. 199–291.

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

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

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

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

Dhurandhar, A. and Dobra, A. 2011. Distribution free bounds for relational classification. *Knowledge and Information Systems*.

Domingos, P. and Richardson, M. 2004. Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML 2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*. 49–54.

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

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

Grimmett, G. and Stirzaker, D. 2001. *Probability and Random Processes*, 3 ed. Oxford.

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

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

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

Jensen, D. and Neville, J. 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In *In Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann, Sydney, Australia, 259–266.

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

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

Kullback, S. and Leibler, R. 1951. On information and sufficiency. *Annals of Mathematical Statistics 22*, 49–86.

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

Malvestuto, F. 1989. A universal table model for categorical databases. *Inf. Sci. 49,* 1-3, 203–223.

Markov, A. 1890. Ob odnom voproce d. i. mendeleeva. In *Zapiski Imperatorskoi Akademii Nauk SP6*. 1–24.

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

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

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

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

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

Papoulis, A. 1991. *Probability, Random Variables and Stochastic Processes*, 3 ed. McGraw-Hill.

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

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

Richardson, M. and Domingos, P. 2006. Markov logic networks. *Mach. Learn. 62,* 1-2, 107–136.

Sachs, L. 1984. *Applied Statistics*. Springer-Verlag.

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

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

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

A.

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 apply the 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,

$$\begin{aligned}
&P[\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] \geq Nt] \\
&= E[I[\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] - Nt \geq 0]] \\
&\leq E[e^{h(\sum_{i=1}^{N} f_i - \sum_{i=1}^{N} E[f_i] - Nt)}] \\
&= e^{-hNt} E[\prod_{i=1}^{N} e^{h(f_i - E[f_i])}] \\
&= e^{-hNt} \prod_{r=1}^{k} E[\prod_{i=g(r)}^{g(r)+m_r-1} e^{h(f_i - E[f_i])}]
\end{aligned} \tag{2}$$

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,

$$E[\prod_{i=g(r)}^{g(r)+m_r-1} Q_i] \leq u \prod_{i=g(r)+1}^{g(r)+m_r-1} w_i \tag{3}$$

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 $v(h_i) = -h_i s_i + \frac{h_i}{b_i - a_i}d_r(\frac{\sum_{j=g(r)}^{i-1} f_j}{i - g(r)} - E[f_i]) + ln(1 - s_i + s_i e^{h_i})$. 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 $\forall i \in \{g(r) + 1, ..., g(r) + m_r - 1\}$,

$$
E[Q_i|Q_{i-1},...,Q_{g(r)}] \leq e^{d_r \delta h + \frac{1}{8}h^2(b_i - a_i)^2} \tag{4}
$$

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 3. Similarly, we can now bound $E[Q_{g(r)}]$,

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

where $l(h) = -hE[f_{g(r)}] + ln(\frac{b_{g(r)} - E[f_{g(r)}]}{b_{g(r)} - a_{g(r)}}e^{ha_{g(r)}} + \frac{E[f_{g(r)}] - a_{g(r)}}{b_{g(r)} - a_{g(r)}}e^{hb_{g(r)}})$. 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,

$$E[Q_{g(r)}] \leq e^{\frac{1}{8}h^2(b_{g(r)} - a_{g(r)})^2} \tag{5}$$

Thus by equations 2, 3, 4 and 5 we have,

$$P[\bar{f} - E[\bar{f}] \geq t] \leq e^{-hNt + \frac{1}{8}h^2 \sum_{i=1}^{N}(b_i - a_i)^2 + \sum_{i=1}^{k}(m_i - 1)\delta d_i h} \tag{6}$$

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 6 we prove the theorem,

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

□