International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.131-144
http://dx.doi.org/10.14257/ijdta.2015.8.3.11

A Dependence Stability Bound based on the VC Dimension for Relational Classification

Xing Wang, Hui He, Bin-Xing Fang and Hong-Li Zhang

Department of Computer Science and Technology, Harbin Institute of Technology, Heilongjiang, P.R.China, 150001
yeahwx@gmail.com; {hehui, bxfang, zhl}@pact518.hit.edu.cn

Abstract

Relational classification (RC) is concerned with the application of statistical learning to relational data. RC models do not have improved stability to smooth the perturbations generated by variations in the correlation between the relational data. Therefore, few studies have attempted to derive a bound and develop a stability learning framework for RC models. To solve this problem, we derive a learning bound with a new measure dependence stability and a limited Vapnik–Chervonenkis (VC) dimension. Based on the learning bound, we then design a stable learning framework that serves as a guideline for the development of new learning algorithms for a broad class of RC models. Applying a Markov logic network on synthesized and real-world datasets, our experimental results demonstrate that our bound can be tight if the RC model has appropriate dependence stability and limited VC dimension and our learning framework increases the stability of RC models while reducing the deviation between empirical risk and true risk.

Keywords: *Dependence Stability, Learning Bound, VC Dimension, Relational Classification*

1. Introduction

Relational data consists of objects and the relationships between these objects are termed as links. Each object has a class label and some attributes. Relational data represented at the individual object and link level as a graph is called a relational data graph [1], wherein the vertices are the objects or attributes and the edges are the links. The left box in Figure 1 shows a toy training data. The data have three objects. Each object has two attributes and a class label.

However, previous works on knowledge transfer learning transfer knowledge in a one-to-one fashion, i.e., only from a single source domain to a single target domain. The knowledge transferred from a single source domain may not be enough to solve new problems. In contrast, Humans are far better than machines as they can learn knowledge from different domains. For example, in scientific innovation, humans get knowledge from multiple disciplines and generate new knowledge to solve new problems. What is missed in machine transfer learning is the ability to create new knowledge from different domains and to transfer pivot knowledge appearing frequently in most of domains.

The learning process of RC models includes structure learning and parameter learning. In some cases the relations among in the objects are explicitly given. In this paper we focus on some cases that the dataset contains implicit relations, e.g. the relation is hidden inside of noisy attribute values. Manually extracting relations by a domain expert is an expensive and time consuming task. To solve the problem, during the structure learning the RC model searches the relational template to capture the relationships of relational training data. For parameter learning, given the relational template, the RC model searches the best parameter for fitting the data. In the process of relational classification, related objects are classified simultaneously. This procedure is common to graphical

ISSN: 2005-4270 IJDTA

models that assume some form of the instance dependence, including Probabilistic Relational Models (PRMs) [2], Markov logic Networks (MLNs) [3], and others. As an example, Figure 1 graphs the learning and classification process of RC models.

Unfortunately, RC models based on maximum likelihood general learning methods (such as, MLN, Relational Dependence Networks (RDNs)[4], *etc.*) are often instable in the strength of dependence between related objects. Ahmed and Neville empirically investigated this case and observed that the dependence do vary significantly throughout the test data. This observation implies that the generalization of RC models is impossible. In contrast, empirical results [6, 7] and a recent statistical analysis suggest that such a generalization is possible, if a single or few examples having small internal correlation and the models have a suitably controlled capacity[8]. These empirical investigation motivate us to control the dependence stability of the RC model.

In this paper, for quantization study on the instable in dependence, we adopt two dependence measures that Dhurandhar proposed [9, 10]. One measure is the number of independent subsets k . Another measure is the dependence strength d [1], which measures the dependence within every subset. Based on the two dependence measures, we first propose a novel dependence stability measure. The measure parameterizes the independent subsets k and the dependent of the RC model d to restrict small changes in the input data. Therefore, the dependence stability enables finer control over the smoothness of the generalization error. We then use the dependence stability and VC dimension to derive a learning bound. By a detailed analysis of the feasibility of this bound, we obtain the conditions under which RC models are learnable, and the necessary criteria for the bound to be tight. Finally, based on the analysis we design a stable learning framework that can be used to develop novel structure and parameter learning algorithms.

We test the stable learning framework empirically. Our experiments on synthesized and real-world datasets demonstrate that: (1) our bound can be tight if the RC model has an appropriate dependence stability and limited VC dimension; (2) our stable learning algorithm simultaneously increases the stability while reducing the deviation between empirical risk and true risk.

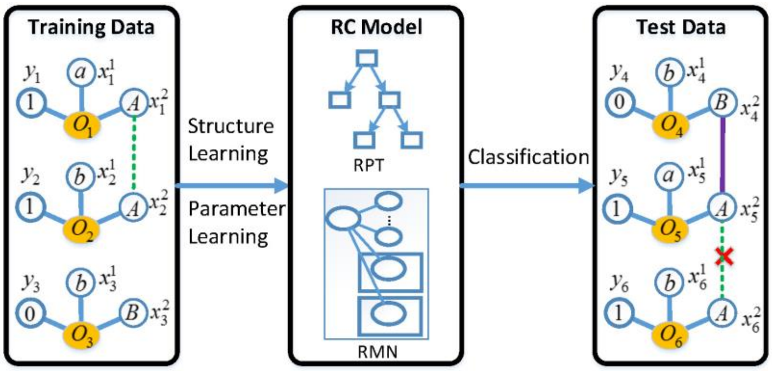

Training Data
RC Model
Test Data
Structure Learning
Parameter Learning
Classification
RPT
RMN
y_1 y_2 y_3 y_4 y_5 y_6
x_1^1 x_1^2 x_2^1 x_2^2 x_3^1 x_3^2 x_4^1 x_4^2 x_5^1 x_5^2 x_6^1 x_6^2
O_1 O_2 O_3 O_4 O_5 O_6
a A b B 1 0

Figure 1. Learning and Classification Process of RC Models

The Training and Test Data have three objects respectively. Each object has two attributes, and a class label. The green dash line represents an implicit relation in the data. The purple solid line represents a relation is determined by the RC model rather than implicit in the data. Note that, the relation that determined by the RC model may not fully fit the implicit relation in test data (the cross on the dash line in the test data means the implicit relation is unfortunately ignored).

2. Preliminaries

In this section, we first set up the learning framework. We then make a detailed introduction to the two dependence measures k and d .

2.1. Set up

As all graphs can be viewed as a hypergraph, the definition of uncertain graph [8] can be extended to uncertain hypergraph directly.

We consider the familiar relational learning setting where the learning algorithm receives a sample of N labeled objects $Z=(z_1,...,z_N)=((x_1,y_1),...,(x_N,x_N,...,y_N))\in(X\times Y)^N$, where X is the attributes space of input objects and the Y is the label set, which is $\{0,1\}$ in classification. In the process of relational classification, related objects are classified simultaneously. We denote a relational classifier by $M:X^N\to\hat{Y}^N$, where the $\hat{Y}^N$ is output set.

The relational classifier has ability to determine the relation between objects (The purple solid line in Figure 1), thereby contributes to the construction of the object dependency graph. Let $G:X^N\to\{e\}_{i=1}^l$ be the edge indicator function, the X^N is input space and the output $\{e\}_{i=1}^l$ is a set. The $e\in\{0,1\}$ indicates that whether there is an edge link two objects in the object dependency graph. When the object dependency graph have N vertexes, the maximal edge number of the graph is $l=N(N-1)/2$. Let $I:\{e\}_{i=1}^l\to k\in[1,N]$ be a independent subset counting function. Function I input linking information, and output the number of independent subset of the dependency graph.

For study the learning bound of RC model, we denote the relational classifier set by $\Theta=\{M_1,M_2,\cdots,M_p\}$, where the p is number of relational classifier in set. Let $L:Y^N\times\hat{Y}^N\to R$ be the loss function, where Y is the label set. The *expected loss* of the particular classifier M be $R(M)=E[L(M(x),y)]$, and the *empirical loss* be $\hat{R}(M)=\frac{1}{N}\sum_{i=1}^N L(y_i,M(x_i))$.

2.2. Dependence Measures

In some cases, the object dependency graph is disconnected, and consists of independent subsets (subgraphs). Each independent subgraph is connected. Based on this observation, Dhurandhar and Dobra proposed two measures to characterize the data relation[9][10]: the number of independent subsets k in the range $[1,N]$ (N is number of object in data), and the dependence strength d [1], which measures the dependence of every subset in the range $[0,1]$.

- The number of independent subsets k capture the subset (subgraphs) property of the object. In general, the more number of independent subset (subgraphs), the more independent in relational data.
- The dependence strength d measures the degree of similarity between the related objects. For relational data, this statistical dependence is called relational autocorrelation.

2.3. Calculation of the Dependence Measures

The number of independent subset (subgraphs) k can be obtained by independent subset counting function (repeating depth-first search on the object dependency graph). For computing d , we adopted the normalized version of the Kullback–Leibler divergence [11].

$$d = \left| \frac{\sum_{i=1}^{k} KL(p_i \parallel q)}{kH_q} \right|$$

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.

Although some relational models assume the autocorrelation is stable throughout the data, many real-world relational datasets exhibit significant variation in k and d with increasing N. We use the expectation and variance of k and d to characterize this variability. Let $E_N[k] = f_k(N)$ and $E_N[d] = f_d(N)$ be the sampling expectation functions of N about k and d, and let $var[k] = g_k(N)$ and $var[d] = g_d(N)$ be the variance functions of N about k and d, respectively. Instead of considering the relational data to describe a temporal process [10, 12], we are able to further study the successive process with these expectation and variance functions. The function can be formulated as $\mathrm{D}:(Y^N, \{e\}_{i=1}^{l}, k) \to d$.

3. Dependence Stability of RC Models

If RC models cannot smooth the fluctuations caused by an increasing volume of sample data, the models are unstable and difficult to generalize. In this section, we define a dependence stability set to restrict the fluctuations in RC models.

To accommodate a variety of loss functions and relational functions, we require the following generic properties. With these properties, the learning bound can be tightened in some case. We recall the definition proposed in literature[12], and revise it accordingly.

Definition 1. A loss function $L: Y^N \times \hat{Y}^N \to R$ is $(B,\lambda)-admissible$ if there exist constants $B < \infty$ and $\lambda < \infty$ such that: (1) for any $y, y' \in Y$ and $\hat{y} \in \hat{Y}$, $|L(y,\hat{y}) - L(y',\hat{y})| \le B$; (2) for any $y \in Y$ and $\hat{y}, \hat{y}' \in \hat{Y}$, $|L(y,\hat{y}) - L(y',\hat{y}')| \le \lambda \|\hat{y} - \hat{y}\|_1$.

Definition 2. An independent subset counting function $I:\{e\}_{i=1}^{l} \to k \in [1,N]$, and a relational autocorrelation function $D:(Y^N, \{e\}_{i=1}^{l}, k) \to R \in [0,1]$ are $(\phi_k, \phi_d)-acceptable$ functions if, for $\forall Y, Y' \in Y^N$, $|I(Y) - I(Y')| \le \phi_k \|(Y - Y')\|_1$ and $|D(Y) - D(Y')| \le \phi_d \|(Y - Y')\|_1$ hold.

Because structure and parameter learning have different processes, better stability measures should be able to characterize both structure and parameter learning. However, the existing stability measure views the learning as separate processes. To solve this problem, we propose a *dependence stability* set to control the two learning processes concurrently.

Definition 3. Let M be a relational classification model from X^n to Y^n. Let the loss function L be $(B,\lambda)-admissible$, and the functions I and D be $(\phi_k, \phi_d)-acceptable$. We say that M has *dependence stability* $\{s_k, e_k, s_d, e_d\}$ if, for any two inputs $x, x' \in X^n$ that differ only at a single coordinate,

$$\begin{cases} \sup\limits_{M \in \Theta} \|I[G(z)] - I[G(z')]\|_1 \le \dfrac{s_k}{\lambda \phi_k \sqrt{N}} \\ \inf\limits_{M \in \Theta} I[G(z)] \ge e_k N (3) \end{cases}$$

hold. For the space limited, we denote the $D(\cdot)$ is the abbreviation of function D.

The set of *dependence stability* has two subsets. The first is defined in formula (2), which includes $\{s_k, e_k\}$. This subset restricts the number of independent subsets k that can be output by the RC model. Thus, the subset can smooth the progress of structure learning. The second subset is defined in formula (3), which includes

$$\begin{cases} \sup\limits_{M \in \Theta} \left\| D(\cdot) - D(\cdot') \right\|_1 \le \dfrac{s_d}{\lambda \phi_d \sqrt{N}} \\ \sup\limits_{M \in \Theta} D[M(X), G(z), I(G(z)] \le e_d \end{cases}$$

$\{s_d, e_d\}$. This subset restricts the dependence strength d of the output of the RC model. This restriction mainly applies to parameter learning.

4. Generalization Bounds

In this section, we use dependence stability to derive PAC Generalization Bounds for RC models. The sufficient conditions for generalization are that the RC models have dependence stability and limited VC dimension.

4.1. Concentration Inequality

It is well known that VC bounds [13] for i.i.d learning are based on Hoeffding-like bounds. However, the Hoeffding bound cannot be directly used on relational data. To overcome this restriction, Dhurandhar and Dobra proposed a distribution-free bound on the generalization error of a non-i.i.d classifier[9][10]. Our study is based on this bound.

Theorem 1. Let N objects $(z_1, ..., z_N)$ be drawn sequentially from relational data with a single-strength dependence parameter d, loss function $\lambda(\cdot,\cdot) \in [0, M]$, and k independent subsets. Assume that $E[\lambda_1] = E[\lambda_2] = \ldots = E[\lambda_N]$, and $\forall i \in 2, \ldots, N$, $E[Z_i \mid Z_{i-1} = z_{i-1}, \ldots, Z_1 = z_1] = \frac{d}{i-1} \sum_{j=1}^{i-1} z_k + (1-d) E[Z_i]$. Then, for $t > \frac{1}{N}(N-k)Md$,

$$P[|Z - \overline{Z}| > t] \le 2e^{\frac{-2[Nt-(N-k)(d+\varepsilon)]^2}{N}}$$

where $\varepsilon = max_i(\varepsilon_i)$ is the error produced by violating the assumptions.

Because the bound uses both the number of independent subsets and the autocorrelation values as a parameter, it will be more useful than other bounds, especially when most of the objects are linked via weak correlation [10]. However, the bound is a temporal value, and k and d are computed from N given relational data.

4.2. Dependence Stability Learning Bounds

In this section, we present a definition of the VC dimension for the self-contained. We then state our main result concerning the learning bound of RC models.

Definition 4. The VC dimension of a hypothesis class C, denoted by VC $dim(C)$, is the cardinality d of the largest set S shattered by C. If all sets S (arbitrarily large) can be shattered by C, then VC $dim(C) = \infty$. Otherwise, VC $dim(C) = max\{d \mid \exists \mid S \mid = d, and \mid growfunction(S) = 2^d \mid\}$.

We now state our main results.

Theorem 2. For any $\delta > 0$, $\delta_k > 0$, $\delta_d > 0$ and $B = \frac{(1-\delta_k)(1-\delta_d)}{\delta+\delta_k\delta_d-\delta_d-\delta_k}$,

$$R(M) \le R(M) + \frac{4}{N}\sqrt{\frac{N}{2}\{\ln 2 + \mathrm{d}_\Theta \ln(\frac{2eN}{\mathrm{d}_\Theta}) - \ln B\}}$$

$$+\frac{4}{N}[N - e_k N + s_k\sqrt{\frac{1}{2\delta_k}}][e_d + s_d\sqrt{\frac{1}{2\delta_d}} + \varepsilon]$$

holds with probability $1-\delta$ over N samples.

Note that, the parameters in the learning bound include extensive measures: the dependence stability ($\{s_k, e_k, s_d, e_d\}$), the complexity of the model (VC dimension d) and data sampling stability (expectation and variance of k and d). Thus, we can make a detailed analyzes of the feasibility of the bound.

We prove Theorem 2 via a series of lemmas. The first lemma establishes the bridge between the variance functions of k, d and the dependence stability.

Lemma 1. If M has dependence stability $\{s_k, e_k, s_d, e_d\}$, the following inequalities hold:

$$g_d(N) \le \frac{s_d^2}{2} \quad \text{and} \quad g_k(N) \le \frac{s_k^2}{2}.$$

The following we extend the standardization symmetrization lemma [14] to non-i.i.d classifiers. The extended lemma replaces the maximum difference between empirical and real risks by the maximum difference between two empirical risks. This replace is useful because the estimation value of classifier is much easier to compute than real risk.

Lemma 2. For any $\varepsilon > 0$, let $v = (N - f_k(N) + \sqrt{\frac{g_k(N)}{\delta_k}})(f_d(N) + \sqrt{\frac{g_d(N)}{\delta_d}} + \varepsilon')$, such that $\frac{1}{N}\left[\frac{1}{2}N\varepsilon - v\right]^2 \ge \frac{1}{2}\ln\frac{(1-\delta_k)(1-\delta_d)}{\frac{1}{2}-\delta_d-\delta_k+\delta_k\delta_d}$, and with probability at least $(1-\delta)$, the following result holds:

$$P(\sup_{M\in\Theta}(R(M) - R(M)) > \varepsilon) \le 2P(\sup_{M\in\Theta}(R'(M) - R(M))) > \frac{\varepsilon}{2})$$

We use the Lemma 2 to build a learning bound. We also add function of expectation and variance of k and d to the learning for characterizing the significant variation in k and d with increasing N (introduced in preliminaries section).

Lemma 3. Let the classification model space is Θ, let $E[k] = f_k(N)$ and $E[d] = f_d(N)$ be the expectation functions of N about k and d, respectively, and $var[k] = g_k(N)$ and $var[d] = g_d(N)$ be the variance functions of N about k and d, respectively. Then we have that

$$P\left(\sup_{M\in\Theta}(\hat{R}'(M) - \hat{R}(M)) > \varepsilon\right) \le \left(\frac{2\mathrm{e}N}{d_\Theta}\right)^{d_\Theta}(1-\delta_k)$$

$$(1-\delta_d)\mathrm{e}^{-\frac{2}{N}[N\varepsilon-(N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}})(f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon')]^2} + \delta_d + \delta_k - \delta_k\delta_d$$

hold.

5. Stable Learning Framework

We want to apply this bound to guide the designing of learning algorithm. In this section, we first analyzes the feasibility of our bound. Then, based on this analysis, we introduce a stable learning framework for RC models.

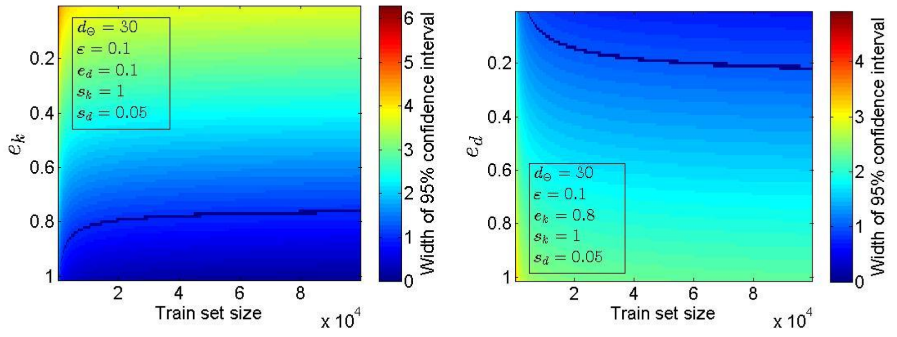
5.1. Feasibility Analysis

The VC bounds ensure that, as the number of sample data increases, the classifier is learnable if and only if the VC dimension of the classifier is limited. However, that the bound is learnable is not sufficient to guarantee it is tight. Thus, we now investigate how the bound varies as the parameters (d_Θ, ε, N) and the dependence stability (s_k, e_k, s_d, e_d) change.

We set $d_\Theta = 30$ for all experiments, thus focusing on the influence of varying the dependence stability of the RC models. We plot the experimental results in Figure 2, and observe the following:

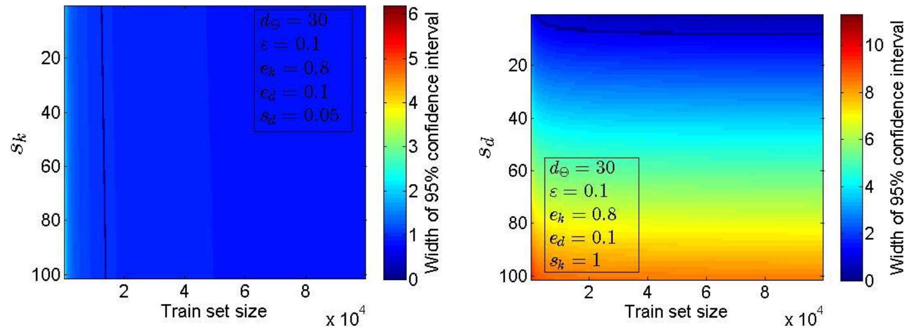
- The bound decreases with increasing N in all sub-figures, especially in Figure 2(c). With the current parameter settings, this trend suggests that the relational model is learnable if the model has a limited VC dimension.
- Figure 2(a), (b) shows that increasing e_k and decreasing e_d leads to a tight bound.
- The bound is insensitive to variations in s_k as shown in Figure 2(c).
- The variation of s_d has a substantial impact on whether the bound is non-trivial. Figure 2(d) shows that, although we set a large e_k and small e_d , the bound is still trivial when s_d is greater than 5.

These observations illustrate that the dependence stability of RC models is an important factor for getting a tight bound. In particular, a large e_k , small e_d and s_d leads to a non-trivial bound. These observations comply with some assumptions about obtaining a tight bound [12]. They assumption that the data exhibits a weak dependence and the predictor exhibits certain complexity and stability properties. In this study, we use a large e_k and small e_d to represent the weak dependence, and use the VC dimension to represent the complexity of models.



(a) Varying e_k and N

(b)Varying e_k and N



(c) Varying s_k and N (d)Varying s_d and N

Figure 2. Each Point is the Average over Four Folds. The Solid Lines Represent a Moving Average over a Five-point Window. The Red Color Represent the $C=1$, the Blue Color Represent the $C=0.2$, the Green Color Represent the $C=0.1$.

5.2. Learning Framework Design

According to above analysis, a stable learning algorithm must ensure a large e_k, small e_d, s_d and limited VC dimension.

- We ensure a large e_k in structure learning. We found the relational template and data determining the number of independent subset of RC models. For example, if all relational data are instances of a relational template, the dependency graph of the model is connected. Thus, the number of independent subsets k is 1. Additionally, when the relational template is determined, changing the parameters of the RC model only influences the dependence within each subset. Thus, in our learning framework, during structure learning, we mainly restrict e_k. Based on this consideration, line 5 to line 7 in Algorithm 1 ensures the e_k is large.
- We restrict e_d and s_d in the parameter learning. In line 8, we add a punishment item d to guarantee dependence stability during parameter learning. The more detail parameter learning process we will introduce in section 8.2.
- We also restrict the complexity of the relational template to indirectly obtain a small VC dimension of the RC models. Note that, the complexity of RC models increases with the number of relational template. A relational template can be regarded as a first order logic formula in MLN or a mode in RDN. This observation is in agreement with preceding studies [15] [16] [17] that restricted the learned relational template in the structure learning process to prevent over-fitting.

Algorithm 1: Stable learning framework

Input: relational data T, e_k, e_d, s_k, s_d, a threshold *minGain*

Output: RC models include a set of *relational template* and parameters

```
1 while no candidate relational schemas added or Gain < minGain do
2         Generate or revise the candidate relational schemas with limited
      complexity
```

```
3    Create the data graph based on current candidate relational schemas and
     relational data
4    Compu the $k$ of data graph by repeated $depth-first$ searches ($by$ a
     independent subset counting function $I$)
5      if $k/N < e_k$ then
6      Heuristic delete some candidate relational schemas that decrease the $k$
7      end
8    Para learning with current candidate relational schemas,
     iterative learning process for the restricted $e_d$, $s_d$, and obtaining an
     optimum fitting value for Gain
9  end
```

The structure learning method in line 2 is already widely applied in many RC models. For instance, the learning algorithm of MLN will predefine the maximum length of the first logic formula, and the RDNs will limit the maximum depth of the tree. Note that, the main difference between our learning framework and existing method is not the method in line 2, but the method of finer-grain control e_k, e_d and s_d that describe in line 4 to 8.

How to use the stable learning framework for a broad class of RC models? Briefly, first we need to add an independent subset counting function I in structure learning process and ensure the number of independent subset of the dependence graph is large enough. Second we have to add the stability punishment item in the optimizing process in parameter learning (the more detail information is described in Section 8.2).

6. Experiments

6.1. Synthetic and Real Datasets

We evaluated the dependence stability using two real datasets. The first contains a classification of webpages from a subset of the WebKB data set, as preprocessed by literature [18]. The processed dataset consists of networks of webpages categorized by course, faculty, project, staff, and student. The pages were collected from four universities, and each page is annotated with word occurrences and links. This preprocessed version of the WebKB data set is relatively small, containing on average 219 pages and 402 links per school.

The second real dataset was the Internet Movie Database (IMDb), downloaded from the alchemy system [19]. The classification task involved identifying 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.

6.2. Dependence Stability Learning

To instantiate a structured predictor with which we can experiment by adjusting the dependence stability, we modify a MLN structure learning algorithm [16] and a variant of Max-Margin Markov Logic network (M3LN) parameter learning method [20].

We modify the structure learning algorithm by mainly restricting e_k, and also restrict the complexity of the relational template indirectly to obtain a small VC dimension of the RC models. We modify the M3LN framework by augmenting the inference objective with a dependence stability regularization term. Though the theory provides guarantees for dependence stability with respect to the 1-norm, in these preliminary experiments, we use a squared 2-norm as the dependence stability term for computational convenience.

The max-margin learning objective is

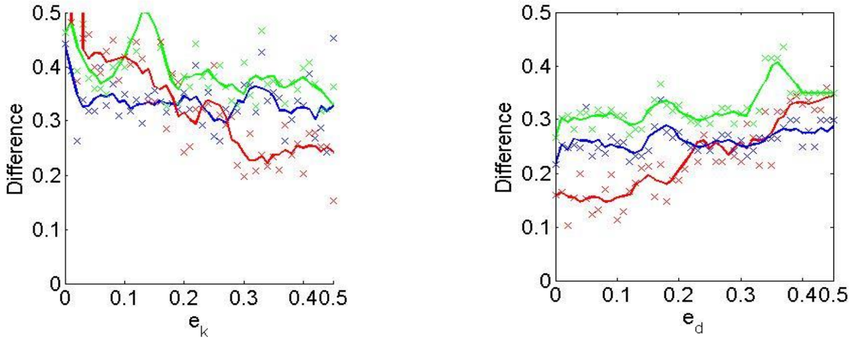
$$\min_{\omega,\xi} \| \omega \|_2^2 + C\xi$$

$$s.t. \forall y' \in Y : \omega[n(x,y) - n(x,y')] - \frac{s_d}{\phi_d \lambda \sqrt{N}} (\|y - y'\|_2^2)] - e_d D(y') \le \xi - \Delta(y,y')$$

The entire learning progress is same as Algorithm 2, which includes structure and parameter learning simultaneously. In each experiment, we apply a variety of slack parameters (C: 0.1, 0.2, 1) and a range of dependence stability parameters (e_k [0, 0.5], e_d [0, 0.5]). To evaluate our predictions, we compute the classification error rates on both the training and test sets. In Figure 3, we plot the difference between the training and test error rates following four-fold cross-validation.

Because the training is conducted on one relatively small network at a time, changes in dependence stability and C can cause spurious jumps in the score. Thus, we plot smoothed curves in addition to the point estimates. We compute the smoothed curves by taking the average over a five-point moving window.

Examining the accuracies reveals that larger values of e_k or small values of e_d tend to decrease the difference between the training and test error rates. These observation are in according with our feasibility analysis in designing a stable algorithm section.



(a) Varying e_k (b) Varying e_d

Figure 3. Each Point is the Average over Four Folds. The Solid Lines Represent a Moving Average over a Five-point Window. The Red Color Represent the $C = 1$, the Blue Color Represent the $C = 0.2$, the Green Color Represent the $C = 0.1$.

7. Conclusion

In this paper, we have derived generalization bounds for RC models. We analyzed the feasibility of these bounds, and identified two sufficient conditions: a limited VC dimension, and a new measure that is specific to RC, the dependence stability. We proposed an experimental estimation method based on our learning bound to better estimate the VC dimension of RC models, and designed a stable learning framework. Our experimental results demonstrate that our bound can be non-trivial if the two conditions are satisfied, and that our stable learning algorithm increases the stability of RC models while reducing the deviation between empirical risk and true risk.

8. Appendex

8.1. A. Proof of Lemma 1

By Definition 3, we have that

$$
\begin{aligned}
\frac{s_d}{\sqrt{N}} \quad & \geq \sup_{M\in\Theta} \lambda\phi_d \left\| M(x) - M(x') \right\|_1 \\
& \geq \sup_{M\in\Theta} \phi_d \, | L[M(x), y] - L[M(x'), y] | \\
& \geq \sup_{M\in\Theta} | D\{L[M(x), y]\} - D\{L[M(x'), y]\} |
\end{aligned}
$$

The function D satisfies bounded differences property. Using Corollary 1 that [5] described, we have that $g_d(N) \leq s_d^2 / 2$. Repeating this process, we can obtain that $g_k(N) \leq s_k^2 / 2$.

8.2. B. Proof of Lemma 2

Because $1\{R(M) - R(M) > \varepsilon\} \cdot 1\{R(M) - R'(M) < \varepsilon / 2\} \leq 1\{R'(M) - R(M) > \varepsilon / 2\}$ is hold for any two dataset. Taking expectations with respect to another sample (the extra data set is usually called 'virtual' or 'ghost' sample) $T^{'} = \{(x_1^{'}, y_1^{'}), \cdots, (x_N^{'}, y_N^{'})\}$, we have that

$$
\begin{aligned}
& 1\{R(M) - R(M) > \varepsilon\} \cdot P\{R(M) - R'(M) < \varepsilon / 2\} \\
& \leq P\{R'(M) - R(M) > \varepsilon / 2\}.
\end{aligned}
$$

Using Theorem 1 we get

$$
\begin{aligned}
& \{1 - \exp\{-\frac{2}{N}[\frac{1}{2}N\varepsilon - (N-k)(d+\varepsilon)]^2\}\} \cdot \\
& 1\{R(M) - R(M) > \varepsilon\} \leq P\{R'(M) - R(M) > \varepsilon / 2\}.
\end{aligned}
$$

Introducing the expectation and variance of k and d, let $v = (N - f_k(N) + \sqrt{\frac{g_k(N)}{\delta_k}})(f_d(N) + \sqrt{\frac{g_d(N)}{\delta_d}} + \varepsilon')$, then $V = \exp\left\{-\frac{2}{N}\left[\frac{1}{2}N\varepsilon - v\right]^2\right\}$ we get

$$
\begin{aligned}
& \left\{1 - (1-\delta_k)(1-\delta_d)V - \delta_d - \delta_k + \delta_k\delta_d\right\} \cdot \\
& 1\{R(M) - \hat{R}(M) > \varepsilon\} \leq P\{\hat{R}'(M) - \hat{R}(M) > \varepsilon / 2\}.
\end{aligned}
$$

Taking expectation with respect to first sample gives the following result,

$$
P\{R(M) - R(M) > \varepsilon\} \leq \frac{P\{R'(M) - R(M) > \varepsilon / 2\}}{\left\{1 - (1-\delta_k)(1-\delta_d)V - \delta_d - \delta_k + \delta_k\delta_d\right\}}
$$

Let $\frac{1}{\left\{1-(1-\delta_k)(1-\delta_d)V-\delta_d-\delta_k+\delta_k\delta_d\right\}} \leq \frac{1}{2}$ and introduce the expectation and variance of k and d, and the N samples are considered to be sampled from data only once, we get the results.

8.3. C. Proof of Lemma 3

If we have a finite set Θ, the union bound immediately yields

$$
P(\exists M \in \Theta) : R(M) - R(M) \geq \varepsilon) \leq \sum_{M\in\Theta} P(R(M) - R(M) \geq \varepsilon) \leq | \Theta | e^{\frac{-2[N\varepsilon - (N-k)(d+\varepsilon)]^2}{N}}
$$

By Lemma 2 and introducing the expectation and variance of k and d, and the N samples are sampled from data only once. By the law of total probability, we have that

$$P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\varepsilon\right)\le|\Theta|(1-\delta_k)$$
$$(1-\delta_d)\mathrm{e}^{-\frac{2}{N}[N\varepsilon-(N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}})(f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon')]^2}$$
$$+\delta_d+\delta_k-\delta_k\delta_d.$$

Replacing $|\Theta|$ with bound of *growth function* $\left(\frac{2\mathrm{e}N}{d_\Theta}\right)^{d_\Theta}$, we have that

which completes the proof.

$$P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\varepsilon\right)\le\left(\frac{2\mathrm{e}N}{d_\Theta}\right)^{d_\Theta}(1-\delta_k)$$
$$(1-\delta_d)\mathrm{e}^{-\frac{2}{N}[N\varepsilon-(N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}})(f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon')]^2}$$
$$+\delta_d+\delta_k-\delta_k\delta_d.$$

8.4. D. Proof of Theorem 2

Using Lemma 3 and Definition 1, we replace the $f_k(N)$, $f_d(N)$, $g_k(N)$ and $g_k(N)$ with e_k, e_d, s_k and s_d, we have that

$$P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\varepsilon\right)\le\left(\frac{2\mathrm{e}N}{d_\Theta}\right)^{d_\Theta}(1-\delta_k)$$
$$(1-\delta_d)\mathrm{e}^{-\frac{2}{N}[N\varepsilon-(N-e_kN+s_k\sqrt{\frac{1}{2\delta_k}})(e_d+s_d\sqrt{\frac{1}{2\delta_d}}+\varepsilon')]^2}+\delta_d+\delta_k-\delta_k\delta_d$$

Let right hand of above inequations equals to δ, and we solve this equation, we have that

$$R(M)\le R(M)+\frac{4}{N}\sqrt{\frac{N}{2}\{\ln 2+\mathrm{d}_\Theta\ln(\frac{2eN}{\mathrm{d}_\Theta})-\ln\delta\}}$$
$$+\frac{4}{N}[N-e_kN+s_k\sqrt{\frac{1}{2\delta_k}}][e_d+s_d\sqrt{\frac{1}{2\delta_d}}+\varepsilon]$$

holds with probability $1-\delta$ over N samples.

Acknowledgements

This work is partially supported by the High-Tech Research and Development Plan of China (No. 2012AA012506); the National Natural Science Foundation of China (Grant No. 61173145, 61472108). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

References

[1] J. Neville, “Statistical models and analysis techniques for learning in relational data”, Ph.D. thesis, University of Massachusetts Amherst, **(2006)**.

[2] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, “Learning probabilistic relational models”, Proceedings of the International Joint Conference on Artificial Intelligence, **(1999)**.

[3] M. Richardson and P. Domingos, “Markov logic networks”, Kluwer Academic Publishers, **(2006)**.

[4] J. Neville and D. Jensen, “Relational dependency networks”, The Journal of Machine Learning Research, vol. 8, no. 9, **(2007)**.

[5] N. Ahmed and Neville, “Network sampling via edge-based node selection with graph induction”, NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks, **(2011)**.

[6] D. Jensen and J. Neville, “Why collective inference improves relational classification”, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining **(2004)**.

[7] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, “Large margin methods for structured and interdependent output variables”, **(2005)**.

[8] R. Xiang and J. Neville, “Relational learning with one network: An asymptotic analysis”, International Conference on Artificial Intelligence and Statistics, **(2011)**.

[9] A. Dhurandhar and A. Dobra, “Distribution-free bounds for relational classification”, Knowledge and information systems, vol. 31, no. 1, **(2012)**.

[10] A. Dhurandhar, “Auto-correlation dependent bounds for relational data”, Proc. of the 11th Workshop on Mining and Learning with Graphs, **(2013)**; Chicago.

[11] S. Kullback and R. A. Leibler, “On information and sufficiency”, The Annals of Mathematical Statistics, vol. 22, no. 1, **(1951).**

[12] B. London, B. Huang, and L. Getoor, “Improved generalization bounds for large-scale structured prediction”, NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks, **(2012)**.

[13] V. Vapnik and A. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities”, Theory of Probability and Its Applications, vol. 16, no. 2, **(1971)**.

[14] V. N. Vapnik, Statistical learning theory, Wiley, **(1998)**.

[15] J. Neville and D. Jensen, “Collective classification with relational dependency networks”, Proceedings of the Second International Workshop on Multi-Relational Data Mining, **(2003)**.

[16] S. Kok and P. Domingos, “Learning the structure of markov logic networks”, Proceedings of the 22nd International Conference on Machine Learning, Association for Computing Machinery, **(2005)**; Bonn, Germany.

[17] L. Mihalkova and T. Huynh, “Mapping and revising markov logic networks for transfer learning”, AAAI, Proceedings of the National Conference on Artificial Intelligence, vol. 1, **(2007)**; Vancouver, BC, Canada.

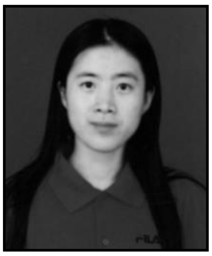
[18] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, “Collective Classification in network data”, AI Magazine, vol. 29, no. 3, **(2008)**.

[19] P. S. S. Kok and M. Richardson, “The alchemy system for statistical relational ai. Dept. of CSE”, Univ. of Washington, **(2007)**.

[20] T. N. Huynh and R. J. Mooney, “Max-margin weight learning for markov logic networks”, Machine Learning and Knowledge Discovery in Databases, vol. 31, no. 57, **(2009)**.

Authors

Xing Wang, he received the B.S. and M.S. degree in computer science from Northwest University, XiAn. China. He is now working towards his Ph.D. degree in computer science at Harbin Institute of Technology. His research interests include computer network, machine learning, and public opinion.



Hui He, she received the B.S., M.S. and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China. Since September 1999, she has been with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, where she became an Associate Professor in October 2007. Her research interests include network computing, network security.

Bin-Xing Fang, he received his M.S. and Ph.D. degrees in computer science from the Tsinghua University and Harbin Institute of Technology of China in 1984 and 1989 respectively. He is currently a member of Chinese Academy of Engineering. His current research interests include information security, information retrieval, and distributed systems.

Hong-Li Zhang, she received her M.S. and Ph.D. in Computer Architecture from the Harbin Institute of Technology on July 1996 and December 1999, respectively. Her research interests are focused in the area of network security, Internet measurement and network computing. She was awarded 3 Ministry Science and Technology Progress awards and published over 50 papers in journals and international conferences.