软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2508−2521 [doi: 10.3724/SP.J.1001.2013.04468]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

关系分类的学习界限研究*

王 星, 方滨兴, 张宏莉, 何 慧, 赵 蕾

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 王星, E-mail: yeahwx@gmail.com

摘 要: 在关系分类模型的学习过程中,目前还没有类似统计学习理论中学习界限的支撑.研究关系分类的学习界限显得尤为重要,为此,提出了一些适用于关系分类模型的学习界限.首先推导出在模型假设空间有限和无限情况下的学习界限.接着提出一个衡量关系模型关联数据能力的复杂性度量——关系维,并证明了该复杂度和关系模型的生长函数之间的关系,得到有限 VC 维和有限关系维下的学习界限.然后分析了该界限可学习和有意义的条件,并对界限的可行性进行了详细的分析.最后分析了基于马尔可夫逻辑网的传统学习界限和关系分类中的学习情况,实验结果表明,所提出的界限能够解释实际关系分类中遇到的一些问题.

关键词: 关系分类;统计关系学习;学习的界限

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 王星,方滨兴,张宏莉,何慧,赵蕾.关系分类的学习界限研究.软件学报,2013,24(11):2508−2521. http://www.jos.org.cn/1000-9825/4468.htm

英文引用格式: Wang X, Fang BX, Zhang HL, He H, Zhao L. Study on relational classification learning bound. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2508−2521 (in Chinese). http://www.jos.org.cn/1000-9825/4468.htm

Study on Relational Classification Learning Bound

WANG Xing, FANG Bin-Xing, ZHANG Hong-Li, HE Hui, ZHAO Lei

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Corresponding author: WANG Xing, E-mail: yeahwx@gmail.com

Abstract: Currently, there is some lack of knowledge about learning bound in relational classification model. In this study, some learning bounds for relational classification are proposed. First, two bounds are deduced for finite and infinite hypothesis space of the relational classification model respectively. Further, a complexity metric called relational dimension is proposed to measure the linking ability of the relational classification model. The relation between the complexity and growth function is proofed, and the learning bound for finite VC dimension and relational dimension is obtained. Afterwards, the condition of learnable, non-trivial, and the feasibility of the bound is analyzed. Finally, the learning progress of relational classification model based on Markov logic network is analyzed with some examples. The experimental result on a real dataset has demonstrated that the proposed bounds are useful in some practical problems.

Key words: relational classification; statistical relational learning; learning bound

关系分类是一种在非独立同分布(non-independent and identically distributed,简称 Non-i.i.d)数据上的分类方法.在关系分类过程中,样本间的关系将直接影响分类的结果.如果关系样本间具有很高的关系自相关值,关系分类的结果将优于在独立同分布数据(independent and identically distributed,简称 i.i.d)上建立的分类模型[1].虽然目前提出了很多关系分类的模型[2−6],这些模型也得到了广泛的应用[2,7−9],但是涉及这些模型的一般学习界限的研究相对较少,因此非常有必要对该问题进行深入的研究.

目前,与关系分类相似的时间序列数据学习问题中,文献[10,11]针对β混合或φ混合过程,使用数据点间的

* 基金项目: 国家自然科学基金(61173145); 国家高技术研究发展计划(863)(2012AA012506)
收稿时间: 2013-02-28; 修改时间: 2013-07-16; 定稿时间: 2013-08-27

依赖强度得出一个界限.文献[12,13]针对 Non-i.i.d 数据泛化了 PAC Bayes 界限,形成了 Chromatic PAC Bayes (CPB)界限.类似于传统的 PAC Bayes 界限,该泛化获得了一个关于随机分类器的训练集一般界限.但由于该界限需要计算数据点的关系依赖图的分数色数,这是一个 NP 难的问题,因此该界限很难在实际中使用.文献[14]扩展了 Hoeffding 类型的概率界限,将其应用于 Non-i.i.d 测试数据中,和 CPB 界限一样,该界限使用关系依赖图的分数色数,因此也难以实际使用.上述研究的基本思想概括起来就是:将数据分成块间独立、块内相关的一些块,并使用块的数量作为参数来计算关系学习模型的界限.

然而,上述研究难以应用于关系分类的环境下,这是因为研究关系分类的学习界限有 3 个问题亟待解决:

(1) 关系分类模型的复杂度需要一个合适的度量.实际上,一些关系模型,如关系马尔可夫随机场(relational Markov random fields,简称 MRFs)、马尔可夫逻辑网(Markov logic network,简称 MLN)与线性预测器类似,其复杂度也应该相近似.

(2) 需要将目前提出的界限扩展到无限训练样本的情况下.当训练样本增长时,其数据之间的关联也随之改变,在这种情况下,支撑界限的概率不等式可能发生变化,因此界限需要进一步扩展.

(3) 学习界限的可行性需要验证.可学习和有意义的学习界限及其存在条件需要进一步的研究[15].

为此,本文针对以上的 3 个问题深入研究了 Non-i.i.d 训练数据上关系分类的学习界限.首先,我们提出一个数据依赖(data-dependent)的学习界限,在已知整体样本的部分统计信息时,建立了关系分类模型的空间有限情况下的学习界限;其次,在分类模型空间无限的情况下,将传统的 VC 维和我们提出的关系复杂性度量用于关系分类的问题中.通过分析参数对界限的影响,得出关系分类可学习以及界限有意义的条件;最后,在真实的实验中验证了该界限的有效性.

本文第 1 节介绍关系数据和一个关系分类在测试集上的界限.第 2 节阐述数据依赖的关系学习的界限.第 3 节首先讨论界限的可行性,然后进行马尔可夫逻辑网的实例分析.第 4 节总结并给出下一步研究方向.

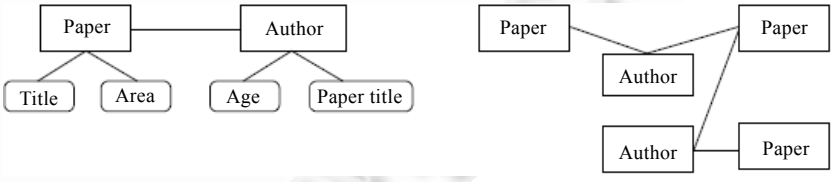
1 关系数据及其测试集界限

1.1 关系数据

关系数据包括对象和对象之间存在的关系即链接.每个对象和链接都有一个类型.相同类型的对象或链接有相同的属性.关系数据的一种表示形式为关系数据图.在该图中,点是对象,边是链接.

另一种关系数据的表示形式是将对象和链接信息存储在表格中的关系数据库管理系统(relational database management systems,简称 RDBMS).一个表格通过特定的类型存储对象或链接信息.列描述关系数据的属性,行则描述个体对象或链接.为了使数据的存储没有冗余,这些表格的信息可以被联合在一起构成一个联合表.

图 1(a)描述了一个关系数据的模式,其具有两个对象类型:Paper 和 Author.图 1(b)为对应的数据图,描述了多个 Paper 和 Author 的链接方式,反映了 Paper 是合著还是单独著作的.



(a) 数据模式 (b) 数据实例

Fig.1 An example of relational data

图 1 关系数据例子

基于关系数据的概率模型是有结构的动态模型,用来处理不确定的关系域.一个模型通过数据图的属性来描述一个联合分布.以图 1 为例,对象 Paper 拥有两个属性:Title 和 Area,对象 Author 也有两个属性:Age 和 Paper

Title.可以通过 Paper 的属性 Title 与 Author 的属性 Paper Title 联合,最终得到一个联合表来描述这些关系数据.例如,一个可能的对象的概率分布为

$$P[Title^1, Area^1, AvgAge^1, ..., Title^6, Area^6, AvgAge^6] = P[Title^1, Area^1, AvgAge^1, Title^2, Area^2, AvgAge^2] \times P[Title^3, Area^3, AvgAge^3, ..., Title^6, Area^6, AvgAge^6],$$

其中,联合对象 1 与联合对象 2 相关,联合对象 3~联合对象 6 相关.

1.2 依赖强度

依赖程度 d 描述的是一个属性与其相关联的属性之间的统计依赖性.在关系数据中,这种统计依赖性被称为关系自相关系数,描述关系数据之间的自相关的程度.下面我们介绍两个关系数据自相关性的计算公式.

自相关系数ρ在连续变量中的求解公式如下:

$$\rho = \frac{\sum_{\{i,j\}\in R}(z_i - \bar{Z})(z_j - \bar{Z})}{\sum_{i\in R} n_i (z_i - \bar{Z})^2} \tag{1}$$

其中,$\forall i \in \{1,...,N\}$,$N$ 是数据集的大小,$z_i \in Z$ 是连续属性的值,$\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} z_i$;$R$ 是数据集中一组相关的数据点.ρ的取值范围为[−1,1].自相关系数ρ在离散数据中求解公式如下:

$$\rho = \frac{\sum_{i=1}^{k} KL(p_i \| q)}{kH_q} \tag{2}$$

其中,k 是数据中独立子集数,p_i 是相应的独立子集中属性值的概率分布,q 是离散属性值的最大熵,KL 为 KL 散度,H_q 是 q 的墒.关系数据的特征是样本之间存在依赖性,或者样本之间由独立性所造成的距离有多远.因此,依赖程度可以描述为 $d=|\rho|$.

1.3 测试集界限

假设有 m 个随机变量 $Z_1,...,Z_m$ 根据联合概率 $P[Z_1,...,Z_m]$变化,如果这些变量存在独立的子集,则可将联合概率进行如下分解:$P[Z_1,...,Z_m]=P[Z_1]P[Z_2|Z_1]...P[Z_m|Z_{m-1},...,Z_1]$,如果假设如下等式成立:

$$\forall i \in \{2,...,m\}\ E[Z_i \mid Z_{i-1} = z_{i-1}, ..., Z_1 = z_1] = d\frac{\sum_{k=1}^{i-1} z_k}{i-1} + (1-d)E[Z_i] \tag{3}$$

则可以通过推导得到 Non-i.i.d 数据的概率不等式[16,17]

$$P(|\bar{Z} - E[\bar{Z}]| \geqslant \varepsilon) \leqslant 2e^{\frac{-2\left(N\varepsilon - \sum_{j=1}^{k}(m_j - 1)\delta d_j\right)^2}{\sum_{j=1}^{N}(b_j - a_j)^2}} \tag{4}$$

成立,对于 $\varepsilon > \frac{\sum_{j=1}^{k}(m_j - 1)\delta d_j}{N}$.其中,$\bar{Z} = \sum_{i=1}^{N}\frac{z_i}{N}$,$\delta = \max_{i,j\in\{1,...,N\}}(b_i - a_j)$.如果$\forall j \in \{1,...,k\}$,$d_j=0$ 并且 $k=N$,那么该不等式可以转化为 i.i.d 数据上的 Hoeffding 不等式:

$$P(|\bar{Z} - E[\bar{Z}]| \geqslant \varepsilon) \leqslant 2e^{\frac{-2N^2\varepsilon^2}{\sum_{j=1}^{N}(b_j - a_j)^2}} \tag{5}$$

给定一个关系数据的依赖性强度系数 d、损失函数$\lambda_i=\lambda(\zeta(x_i),y_i)$,$\forall i \in \{1,...,N\}$,$\lambda$的取值范围为[0,a],$k$ 个独立子集.并假设 $E[\lambda_1]=E[\lambda_2]=...=E[\lambda_N]$,那么从公式(4)可以得知,当 $\varepsilon > \frac{(N-k)ad}{N}$ 时有:

$$P[|HE - GE| \geqslant \varepsilon] \leqslant 2e^{\frac{-2(N\varepsilon - (N-k)ad)^2}{Na^2}} \tag{6}$$

2 数据依赖的学习界限

这一节,我们将把统计学习中经典的 VC 维界限扩展到关系分类问题中,扩展中有 3 个问题需要解决:

(1) 确定 VC 维是否能适用于关系分类模型.很多流行的机器学习模型的 VC 维已经得到了广泛的研究,

但是关系分类模型还没有得到很好的研究.实际上,VC 维衡量了给定函数集合的复杂性,其定义与数据的独立同分布假设是否成立没有关系.我们发现,有些关系分类模型具有有限的 VC 维,具体的实例将在第 3.2 节进行详细介绍.

(2) 公式(6)只适用于给定数量的测试样本,不能适用于逐渐增加和分布改变的训练样本.这是因为当样本给定时,其关系自相关系数 d 以及独立子集个数 k 也随之确定,那么公式(6)也将确定.而真实世界中的关系数据是采样而来的,其 d 和 k 会随着 N 的增加而改变,那么公式(6)也会随之改变.为了描述这种改变,我们引入 k 和 d 关于 N 的期望函数 $E_N[k]=f_k(N),E_N[d]=f_d(N)$及其相应的方差函数 $var[k]=g_k(N)$ 和 $var[d]=g_d(N)$,以便深入研究在关系数据波动的情况下学习的界限.

(3) 有必要衡量关系分类模型关联数据的能力及其关联子模型的复杂度.当数据中关系较为普遍和复杂时,关系模型相对于统计学习模型能够更好地表示这些数据间存在的关联,因此能够更好地拟合关系数据.为了更好地刻画该能力,我们提出一个新的度量——关系维,来衡量关系分类模型关联子模型的关联能力及其复杂度.

接下来,我们首先推导有限的模型空间下的学习界限;其次,提出一个关系复杂度的度量并给出其生长函数的界;最后,给出无限模型空间下的学习界限.

2.1 有限的模型空间下的界限

已知关系分类训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$,它是从关系数据联合表的联合概率分布 $P(X,Y)$非独立但同分布关系采样获得的.设关系分类的模型集合$\Theta=\{M_1,M_2,\ldots,M_{|\Theta|}\}$,损失函数 L 是 0-1 损失.关于Θ的期望风险和经验风险分别是 $R(M)=E[L(y,M(x))],\hat{R}(M)=\dfrac{1}{N}\sum_{i=1}^{N}L(y_i,M(x_i))$.

下面我们引入 k 和 d 关于 N 的期望函数 $E_N[k]=f_k(N),E_N[d]=f_d(N)$及其相应的方差函数 $var[k]=g_k(N)$和 $var[d]=g_d(N)$,得到一个有限空间下的学习的界限.

定理 1(假设空间有限下的学习界限). 设模型空间是一个有限的模型集合$\Theta=\{M_1,M_2,\ldots,M_{|\Theta|}\}$,已知 k 和 d 的期望值分别是关于 N 的函数,即 $E_N[k]=f_k(N),E[d]=f_d(N)$,相应的方差为 $var[k]=g_k(N),var[d]=g_d(N)$,则

$$R(M)\leqslant\hat{R}(M)+\frac{1}{N}\sqrt{\frac{N}{2}\ln\frac{|\Theta|}{\delta}}+\frac{1}{N}\left[N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}}\right]\left[f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon'\right]$$

在 N 个采样的关系数据上以$(1-\delta)(1-\delta_k)(1-\delta_d)$的概率成立.

证明:由于 $\mathfrak{M}=\{M_1,M_2,...,M_{|\mathfrak{M}|}\}$ 是一个有限的集合,所以,

$$P(\exists M\in\Theta:R(M)-\hat{R}(M)\geqslant\varepsilon)\leqslant\sum_{M\in\Theta}P(R(M)-\hat{R}(M)\geqslant\varepsilon)\leqslant|\Theta|\,\mathrm{e}^{\frac{-2[N\varepsilon-(N-k)(d+\varepsilon')]^2}{N}}$$ (见文献[16−18]).

其中,ε'为概率不等式的误差值.

令$|\Theta|\exp\left\{-\dfrac{2}{N}[N\varepsilon-(N-k)(d+\varepsilon')]^2\right\}=\delta$, 将$\varepsilon$代入,则

$$R(M)\leqslant\hat{R}(M)+\frac{1}{N}\sqrt{\frac{N}{2}\ln\frac{|\Theta|}{\delta}}+\frac{(N-k)(d+\varepsilon')}{N}$$

以 $1-\delta$的概率成立.

将 k 和 d 的期望与方差函数代入可得:

$$R(M)\leqslant\hat{R}(M)+\frac{1}{N}\sqrt{\frac{N}{2}\ln\frac{|\Theta|}{\delta}}+\frac{1}{N}\left[N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}}\right]\left[f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon'\right]$$

以$(1-\delta)(1-\delta_k)(1-\delta_d)$的概率成立.

2.2 关系复杂度

在关系数据上建立的关系模型与传统 i.i.d 数据上的分类模型相比,能够更好地拟合数据间存在的关系.这

可以理解为关系模型蕴含了两个方面的能力:一是与传统 i.i.d 数据上的分类模型一致的分类能力,二是拥有将数据关联起来,然后集合分类的能力.基于这种考虑,为了细粒度的分析关系模型的分类能力,我们将关系分类模型集合Θ划分为处理独立同分布数据的子模型Θ_I和用于关联数据的子模型Θ_R.

关系模型集合Θ的复杂性度量,我们仍然使用VC维d_Θ来衡量,这是因为一些关系模型如关系马尔可夫随机场(relational Markov random fields,简称 MRFs)、马尔可夫逻辑网(Markov logic network,简称 MLN)与线性预测器类似,其复杂度也应该相近似.Θ_I处理的是 i.i.d 数据,因此我们仍然使用 VC 维 d_I来衡量其复杂度.子模型Θ_R的能力是关联数据,目前还缺少衡量其能力及复杂度的合适的度量.为此,我们提出一个新的度量——关系维 d_R来衡量关系分类模型的关联数据能力及其复杂度,并证明了该复杂度和关系模型的生长函数之间的关系.

定义 1. 对于给定的 N 个关系数据和关系分类模型集合Θ,关系生长函数 $S_R(N)$是该模型集合能够表示的关系数据连接方式的最大数量:

$$S_R(N)=\max_{\{x_1,\ldots,x_N\}\subset X}|\{(\Theta_R(x_1,x_2),\Theta_R(x_1,x_3),\ldots,\Theta_R(x_i,x_j)):\forall i,j\in\{1,\ldots,N\},i\neq j,\Theta_R\in\Theta\}|,$$

其中,Θ_R 是关系分类器Θ的一个抽象子模型,拥有关联样本的能力.$\Theta_R(x_i,x_j)$=1指示样本x_i和x_j有关联, $\Theta_R(x_i,x_j)$=0 指示这两个样本没有关联.

定义 2. 给定关系分类模型集合Θ,其关系维 $V_R(\Theta)=\max\{N:S_R(N)=2^{N(N-2)/2}\}$,也就是能够将 N 个关系数据的 $N(N-1)/2$ 个连接方式都表示出来的最大数据个数.

定理 2(关系生长函数的界). 给定关系分类模型集合Θ,其关系维 $V_R(\Theta)=d_R$,那么对于任意的 N 个数据,有:

$$\Lambda_\Theta(N)\leqslant\left[\frac{\mathrm{e}N(N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}.$$

证明:类似于子模型Θ_I 的生长函数 $S_I(N)$的输入为数据的个数 N,可以视这 N 个数据间最大的连接边数 $N(N-1)/2$ 为子模型Θ_R 的输入个数,进而得到关系生长函数:

$$S_R(N)=S_I(N(N-1)/2)\leqslant\left[\frac{\mathrm{e}\frac{N(N-1)}{2}}{\frac{d_R(d_R-1)}{2}}\right]^{\frac{d_R(d_R-1)}{2}}=\left[\frac{\mathrm{e}N(N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}.\quad\square$$

引理 1(模型生长函数的界). 给定关系分类模型集合Θ,其整体模型集合的 VC 维存在并且等于 d_Θ.模型集合Θ的生长函数为 $S_\Theta(N)$.用于处理独立同分布数据的子模型Θ_I的 VC 维为 d_I,其生长函数为 $S_I(N)$.关联数据的子模型Θ_R 的关系维为 d_R,其生长函数为 $S_R(N)$.假设一个被子模型Θ_I 正确分类的数据只能将与其连接的错误分类数据纠正为与其相同的类别;关联不会损害Θ_I 子模型的分类能力;模型Θ_R 所能表示的任意一种数据连接方式,能够正确纠正Θ_I未能正确分类的数据个数最大为 $C(N)$,那么有:

(1) 当 $N\leqslant d_I$时,$S_\Theta(N)=2^N$;

(2) 当 $N>d_I, N(N-1)/2\leqslant d_R(d_R-1)/2$ 时,$S_\Theta(N)=2^N$;

(3) 当 $N>d_I, N(N-1)/2>d_R(d_R-1)/2, 2^N\leqslant S_R(N)C(N)$ 时,$S_\Theta(N)=2^N$;

(4) 当 $N>d_I, N(N-1)/2>d_R(d_R-1)/2, 2^N>S_R(N)C(N)$ 时,

$$S_\Theta(N)\leqslant S_I(N)+S_R(N)C(N)=\left(\frac{\mathrm{e}N}{d_I}\right)^{d_I}+C(N)\left[\frac{\mathrm{e}N(N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}.$$

证明:对于 N 个关系数据,当 $N>d_I$ 时,子模型Θ_I可以表示的类别指派的数量范围是[0,$S_I(N)$],那么其未能表示的指派数量范围是[2^N–$S_I(N)$,2^N].

(1) 当 $N\leqslant d_I$ 时,子模型Θ_I已经能够正确地表示 N 个关系数据样本的 2^N 个类别指派可能;又因为假设连接不会损害Θ_I子模型的分类能力,因此无论Θ_R 能够表示多少连接,整体模型的生长函数值都为 2^N.

(2),(3) 当 $N>d_I$ 时,子模型Θ_I已经不能完全正确地表示 N 个关系数据样本的 2^N 个类别指派可能.这时需要以子模型Θ_R 的能力,将所有指派正确地表示.当Θ_R 表示的连接方式的数量大于关系维,即当 $N(N-1)/2\leqslant$

$d_R(d_R-1)/2$ 时,无论Θ_I的最大未能表示的类别指派数是否大于Θ_R的最大能纠正的数量,其整体模型还是能将N个样本的2^N个可能都正确地表示出来,因此其生长函数值为2^N.当$N(N-1)/2>d_R(d_R-1)/2$,且Θ_I的最大未能表示的类别指派数2^N小于Θ_R的最大纠正数$S_R(N)C(N)$时,其生长函数的值为2^N.

(4) 当 $N>d_I, N(N-1)/2>d_R(d_R-1)/2, 2^N>S_R(N)C(N)$ 时,整体模型已经不能将 N 个样本的 2^N 个可能都表示正确.但是当假设成立时,其生长函数的界为Θ_I所能表示的最大类别指派数和Θ_R所能表示的最大纠正数之和,即 $S_I(N)+S_R(N)C(N)$.

综上所述,引理 1 得证. □

2.3 无限样本下的界限

在模型的空间无限的情况下,需要将其映射到有限的空间中,才能得到有效的界限.为此,我们扩展一般的对称性引理到 non-i.i.d 的数据中,然后考察在采样的数据中产生的有限空间随数据量的增长情况.

引理 2(关系模型对称性引理). 对于任意的$\varepsilon>0$,且

$$\frac{1}{N}\left\{\frac{1}{2}N\varepsilon-\left[N-f(N)-\sqrt{g_k(N)/\delta_k}\right]\left[g(N)-\sqrt{g_d(N)/\delta_d}+\varepsilon'\right]\right\}^2\geqslant-2\ln\frac{1}{2},$$

有: $P\left(\sup_{M\in\Theta}(R(M)-\hat{R}(M))>\varepsilon\right)\leqslant 2P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\frac{\varepsilon}{2}\right)$以$(1-\delta_k)(1-\delta_d)$的概率成立.

证明: $1\{R(M)-\hat{R}(M)>\varepsilon\}1\{R(M)-\hat{R}'(M)<\varepsilon/2\}\leqslant 1\{\hat{R}'(M)-\hat{R}(M)>\varepsilon/2\}$.

在不等式两边先对 $T'=\{(x_1',y_1'),(x_2',y_2'),...,(x_N',y_N')\}$ 求期望得到:

$$1\{R(M)-\hat{R}(M)>\varepsilon\}P\{R(M)-\hat{R}'(M)<\varepsilon/2\}\leqslant P\{\hat{R}'(M)-\hat{R}(M)>\varepsilon/2\},$$

$$\left\{1-\exp\left\{-\frac{2}{N}\left[\frac{1}{2}N\varepsilon-(N-k)(d+\varepsilon')\right]^2\right\}\right\}1\{R(M)-\hat{R}(M)>\varepsilon\}\leqslant P\{\hat{R}'(M)-\hat{R}(M)>\varepsilon/2\}.$$

两边对 T 求期望得到:

$$\left\{1-\exp\left\{-\frac{2}{N}\left[\frac{1}{2}N\varepsilon-(N-k)(d+\varepsilon')\right]^2\right\}\right\}P\{R(M)-\hat{R}(M)>\varepsilon\}\leqslant P\{\hat{R}'(M)-\hat{R}(M)>\varepsilon/2\},$$

$$P\{R(M)-\hat{R}(M)>\varepsilon\}\leqslant\frac{1}{1-\exp\left\{-\frac{2}{N}\left[\frac{1}{2}N\varepsilon-(N-k)(d+\varepsilon')\right]^2\right\}}P\{\hat{R}'(M)-\hat{R}(M)>\varepsilon/2\}.$$

若 $\frac{1}{1-\mathrm{e}^{\frac{-2}{N}\left[\frac{1}{2}N\varepsilon-(N-k)(d+\varepsilon')\right]^2}}\leqslant\frac{1}{2}$,那么得到:

$$\frac{1}{N}\left[\frac{1}{2}N\varepsilon-(N-k)(d+\varepsilon')\right]^2\geqslant-2\ln\frac{1}{2}.$$

将 k 和 d 的期望值函数代入,得到条件:

$$\frac{1}{N}\left\{\frac{1}{2}N\varepsilon-\left[N-f(N)-\sqrt{\frac{g_k(N)}{\delta_k}}\right]\left[g(N)-\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon'\right]\right\}^2\geqslant-2\ln\frac{1}{2}.$$ □

定理 3(关系模型有有限 VC 维时学习的界限). 对于任意$\delta>0,\delta_k>0,\delta_d>0$,假设关系分类模型$\Theta$的 VC 维存在并且等于 d_Θ,且数据间的关联关系不受关系模型的影响,那么,

$$R(M)\leqslant\hat{R}(M)+\frac{4}{N}\sqrt{\frac{N}{2}\left\{\ln 2+d_\Theta\ln\left(\frac{2\mathrm{e}N}{d_\Theta}\right)-\ln\delta\right\}}+\frac{4}{N}\left[N-f_k(N)+\sqrt{\frac{g_k(N)}{\delta_k}}\right]\left[f_d(N)+\sqrt{\frac{g_d(N)}{\delta_d}}+\varepsilon'\right]$$

以$(1-\delta)(1-\delta_k)(1-\delta_d)$的概率成立.

证明:通过引理 1,能够得到:

$$P\left(\sup_{M\in\Theta}(R(M)-\hat{R}(M))>\varepsilon\right)\leqslant 2P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\frac{\varepsilon}{2}\right)$$
$$\leqslant 2S_{\Theta}(2N)P\left[\hat{R}'(M)-\hat{R}(M)>\frac{\varepsilon}{2}\right]$$
$$\leqslant 2\left(\frac{2eN}{d_{\Theta}}\right)^{d_{\Theta}}e^{\frac{-2[N\varepsilon/4-(N-k)(d+\varepsilon')]^2}{N}}.$$

将 k 和 d 的期望值函数代入,得到结果. □

定理 4(关系模型有有限 VC 维和有限关系维时学习的界限). 对于任意δ>0,δ_k>0,δ_d>0,关系分类模型集合Θ,假设整体模型集合的 VC 维存在并且等于 d_{Θ},模型集合Θ的生长函数为 $S_{\Theta}(N)$.用于处理独立同分布数据的子模型Θ_I的 VC 维为 d_I,其生长函数为 $S_I(N)$.用于关联数据的子模型Θ_R的关系维为 d_R,其生长函数为 $S_R(N)$.如果引理 1 的假设成立,那么,当 $N>d_I, N(N-1)/2>d_R(d_R-1)/2, 2^N>S_R(N)C(N)$ 时,

$$R(M)\leqslant\hat{R}(M)+\frac{4}{N}\sqrt{\frac{N}{2}\ln 2\left\{\left(\frac{2eN}{d_I}\right)^{d_I}+C(2N)\left[\frac{2eN(2N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}\right\}-\frac{N}{2}\ln\delta+\frac{4}{N}(N-k)(d+\varepsilon')}$$

以$(1-\delta)(1-\delta_k)(1-\delta_d)$的概率成立.

证明:通过引理 1,能够得到:

当 $N>d_I, N(N-1)/2>d_R(d_R-1)/2, 2^N>S_R(N)C(N)$ 时,

$$P\left(\sup_{M\in\Theta}(R(M)-\hat{R}(M))>\varepsilon\right)\leqslant 2P\left(\sup_{M\in\Theta}(\hat{R}'(M)-\hat{R}(M))>\frac{\varepsilon}{2}\right)$$
$$\leqslant 2S_{\Theta}(2N)P\left[\hat{R}'(M)-\hat{R}(M)>\frac{\varepsilon}{2}\right]$$
$$\leqslant 2\left\{\left(\frac{2eN}{d_I}\right)^{d_I}+C(2N)\left[\frac{2eN(2N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}\right\}e^{\frac{-2[N\varepsilon/4-(N-k)(d+\varepsilon')]^2}{N}}.$$

从中解得:

$$\varepsilon=\frac{4}{N}\sqrt{\frac{N}{2}\ln 2\left\{\left(\frac{2eN}{d_I}\right)^{d_I}+C(2N)\left[\frac{2eN(2N-1)}{d_R(d_R-1)}\right]^{\frac{d_R(d_R-1)}{2}}\right\}-\frac{N}{2}\ln\delta+\frac{4}{N}(N-k)(d+\varepsilon')}.$$

将 k 和 d 的期望值函数代入,得到结果. □

3 界限分析与应用实例

这一节我们首先对上一节所推导的界限进行可行性分析,从中得出一些关系分类模型可学习和有意义的条件;然后,将该界限用于一个具体的统计关系模型马尔可夫逻辑网.相关的实例表明:该界限能够很好地解释基于马尔可夫逻辑网的传统分类和集合分类的过程.

3.1 界限可行性分析

3.1.1 关系分类模型可学习的条件

随着关系数据的增长,关系学习模型的平均误差应该逐渐收敛于期望误差,这表明模型是可学习的.但是根据定理 3,该界限受到 4 个参数的影响,而传统的分类模型的界限只受到数据量 N 的影响.我们归纳得到以下的可学习条件:

$$\begin{cases}\dfrac{2}{N}\left\{\dfrac{N\varepsilon}{4}-\psi\right\}^2>\tau\\ \dfrac{1}{N}\left\{\dfrac{1}{2}N\varepsilon-\psi\right\}^2\geqslant-2\ln\dfrac{1}{2}\\ \tau>\ln\delta\end{cases}\tag{7}$$

其中,$\psi=\left[N-f_k(N)+\sqrt{g_k(N)/\delta_k}\right]\left[f_d(N)+\sqrt{g_d(N)/\delta_d}+\varepsilon'\right]$,$\tau=\ln 2+d_\Theta\ln 2eN-d_\Theta\ln d_\Theta$.

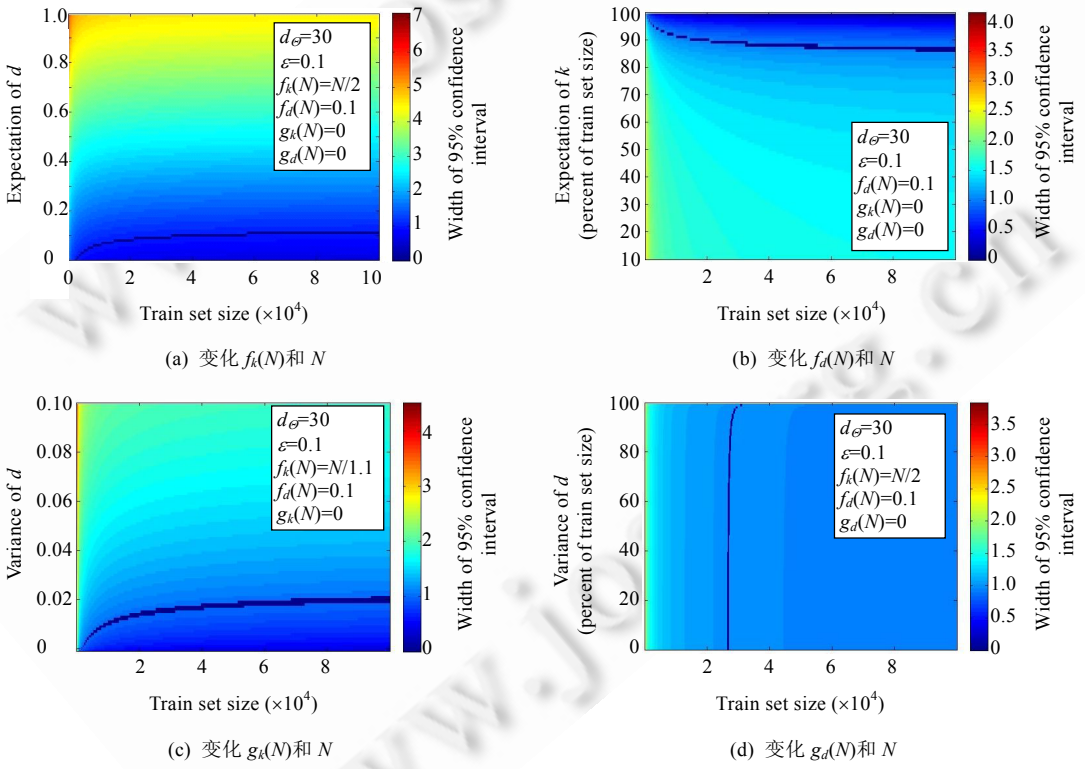
当 $f_k(N)=1$,$f_d(N)=N$,$g_k(N)=N$,$g_d(N)=N$ 时,可以获得最大的界限:

$$R(M)\leqslant\hat{R}(M)+\frac{4}{N}\sqrt{\frac{N}{2}\left\{\ln 2+d_\Theta\ln\left(\frac{2eN}{d_\Theta}\right)-\ln\delta\right\}}+\frac{4}{N}\left[N-1+\sqrt{\frac{N}{\delta_k}}\right]\left[1+\sqrt{\frac{1}{\delta_d}}+\varepsilon'\right].$$

该不等式的最后一项是随着 N 的增长单调递减的,因此当 N 趋向于无穷大时,界限趋向于 0.这表明在这样的条件下,该模型是可学习的.

3.1.2 界限有意义的相关分析

通过以上的分析,关系分类模型在某些情况下是可学习的,但是,界限是可学习的并不能保证学习是有意义的(非平凡的,界限应该小于 1).接下来,我们专注于调查随着某些关系数据采样统计信息的变化对定理 3 提出的界限的影响,如图 2 所示.在图 2 中,X 轴为训练样本数量,Y 轴为统计信息,颜色的深浅表示界限的 95%置信区间的宽度,其中,黑色的线表示宽度为 1 的情况.在以下所有的实验中,我们设定 d_Θ=30.



(a) 变化 $f_k(N)$和 N (b) 变化 $f_d(N)$和 N

(c) 变化 $g_k(N)$和 N (d) 变化 $g_d(N)$和 N

Fig.2 Bounds variation with statistical information of sampling data

图 2 随着采样数据统计信息的变化,界限的变化

从图 2 中我们可以观察到以下的一些影响:

- 当样本数量 N 增加时,在所有图中置信区间的宽度都是减少的,特别是在图 2(c)中.在当前的参数设置

下,这样的趋势说明了如果模型有有限的 VC 维,那么关系模型是可学习的.

- 图 2 中,黑色的线表示 95%置信区间的宽度为 1 的情况.显然,界限有意义的区域(置信区间宽度小于 1 的区域)小于无意义的区域(置信区间宽度大于 1 的区域),并且这样的现象与这些参数值的变化有极大的关系.
- 图 2(a)、图 2(b)表明,增加的 $f_k(N)$和减少的 $f_d(N)$将导致较紧的界限,如同定理 1 所描述的.如果关系数据的 $f_k(N)$和 $f_d(N)$表现出较小的改变速率,那么也会得到一个较紧的界限.
- 当 $f_k(N)$和 $f_d(N)$有较小的改变速率时,该界限对 $f_d(N)$的变化不敏感,如图 2(c)所示.
- $g_d(N)$的变化对界限是否有意义有决定性的影响.图 2(d)显示:当 $g_d(N)\geqslant 0.06$ 时,尽管我们设置了一个较小的 $f_k(N)$和 $f_d(N)$,该界限仍然是无意义的.

当样本数量 N 趋于无穷大时,该界限的极限为 $\lim_{N\to\infty}\frac{4}{N}\left[N-f_k(N)+\sqrt{g_k(N)/\delta_k}\right]\left[f_d(N)+\sqrt{g_d(N)/\delta_d}+\varepsilon'\right]$.

通常情况下,关系数据的参数值取值为 $\delta_d<0.05$,$\varepsilon'=0.1$ 和 $\delta_k=0.05$,这时,界限是否有意义将会极大地受到参数值的影响.

以上的这些观察与文献[15]所提出的界限的稳定性条件一致,该文献假设:如果关系分类数据有弱依赖性,并且分类模型具有一定的稳定性和复杂度,那么所获得的界限是可学习的.本文提出的界限采用较小的关系自相关系数 d 和确定的 VC 维,与其假设一致.

通过以上的讨论,我们总结定理 3 所提出的界限的可行性如下:

- 当关系分类模型有有限的 VC 维,并且不等式(6)成立时,模型是可学习的;
- 有限的 VC 维不能保证界限是有意义的,并且有意义界限的条件很严格;
- 数据采样的稳定性对于关系分类模型的学习界限是一个非常关键的性质,特别是与关系自相关值有关的采样统计信息和数据中独立子集的数量.

3.2 马尔可夫逻辑网实例分析

这一节我们首先介绍一个能够用于关系分类的统计关系学习模型马尔可夫逻辑网,然后分析基于马尔可夫逻辑网的传统学习和关系分类的情况.

3.2.1 马尔可夫逻辑网

马尔可夫逻辑网结合了一阶逻辑和 Markov 网,其基本思想是将一阶知识库的一些硬限制进行软化:当一个世界违反知识库中的一个公式时,其发生的概率较小,但未必为 0.一个世界违反的公式越少,其发生的概率越大.用公式的权值来表示公式限制强度的大小,权值越高,满足该公式的世界发生的概率与不满足该公式的世界发生的概率之间的差就越大.基于这样的基本思想,MLN 定义如下:

定义 3. 马尔可夫逻辑网 L 是二元组(F_i,ω_i)集合,其中,F_i 是一阶逻辑公式,ω_i 是实数.与有限常数集 $C=\{c_1, c_2,\ldots,c_{|C|}\}$一起定义的马尔可夫逻辑网如下:

(1) L 中的每一个闭谓词对应着 $M_{L,C}$ 中的一个二值节点.若闭谓词为真,则节点值为 1;否则为 0.

(2) L 中的一个公式 F_i 可能的闭公式对应着 $M_{L,C}$ 中的一个特征.如果闭公式为真,则其为 1;否则为 0.特征权重ω_i 对应 L 中的 F_i.

该定义可以看作构建马尔可夫网络的模板.如果常数集不同,它会产生大小不同的网络,但是所有关于结构和参数的规则都是确定的(比如一个公式的所有闭公式都有相同的权重).我们把它称作基本马尔可夫网,以示与一阶马尔可夫逻辑网的区别.对于一个可能世界 x,其基本马尔可夫逻辑网概率分布为

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right)=\frac{1}{Z}\prod_i \varphi_i(x_{\{i\}})^{n_i(x)} \tag{8}$$

其中,$n_i(x)$是 F_i 在 x 中所有取真值的基本规则的数量,而 $X_{\{i\}}$是 F_i 中谓词的状态(真值),且有 $\varphi_i(x_{\{i\}})=\mathrm{e}^{w_i}$.

直观上看,马尔可夫逻辑网就是每个准则都有权重的一个一阶逻辑知识库,是构建马尔可夫逻辑网的模板.从概率的视角来看,马尔可夫逻辑网提供一种简洁的语言来定义大型 Markov 网,能够灵活地、模块化地与大量

知识合并.从一阶逻辑的视角来看,马尔可夫逻辑网提供了健全地处理不确定性、容许有瑕疵甚至矛盾的知识库,降低了脆弱性.有许多统计关系学习领域的重要任务,如集合分类、连接预测、连接聚合、社会网络建模和对象识别,都自然而然地成为运用马尔可夫逻辑网推理和学习的实例[18,19].

3.2.2 传统分类问题

通过构造合适的谓词和子句,MLN 能够处理传统的分类问题[20].首先定义特征谓词 feature_i(row,value_i)和类别谓词 class(row,value_c),用以涵盖特征矩阵和分类标签的基本信息.其中,row 为特征矩阵的行号,value_i 中的 i 为特征矩阵的列号,value_c 为类别标签.接着,将谓词组成分类用的一阶逻辑子句.文献[20]详细分析了可能的分类子句形式后得出经验的结论,使用如下的子句集能够很好地描述分类所需的逻辑关系,从而达到较高的分类准确度:

feature_i(row,value_i!)

feature_j(row,value_j!)

…

class(row,value_c!)

class(row,+value_i)^feature_i(row,+value_i)

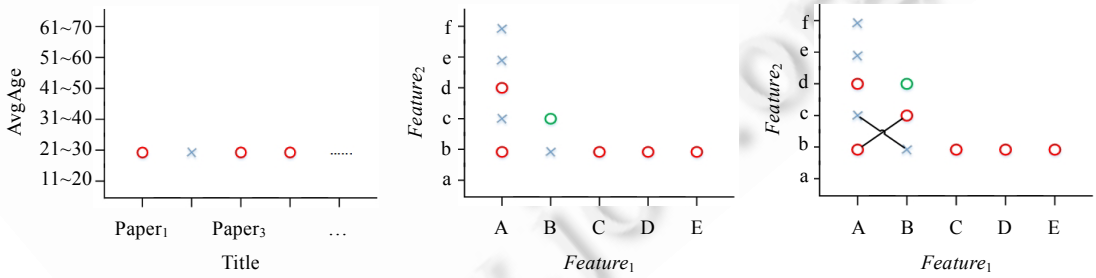
class(row,+value_j)^feature_i(row,+value_j)

class(row,+value_k)^feature_i(row,+value_k)

…

其中,操作符“!”表示 row 有且只有 1 个 value_i 的值.操作“+”表示该值可以被替换为数据中任意一个符合该常量类型的值,使得以上的每一个子句都成为一个子句集.该分类子句集*class(row,+value_i)^*feature_i(row,+value_i)的个数随着样本特征的数目及其取值可能的增加而增加.其输入为一个样本的所有特征值,输出一个总的判断后的输出.

上述字句集合的 VC 维实际上与特征的取值范围有关.例如,特征 Paper 的 Title 属性在图 1 中的取值范围是无穷的,因此,最大能够分开的集合可以简单地将数据点沿着无限特征轴放置.图 3(a)这样的模型的 VC 维是无穷大的(在图 3 中,X 轴和 Y 轴是两个特征,黑色的圆圈表示该实例的标签为正,黑色的十字表示该实例的标签为负,浅色的圆圈则表示该实例被误分).但在大多数情况下,MLN 要求特征的取值范围是有限的.在这样的情况下,以上子句集的 VC 维将是有限的.例如图 3(b),假设有 2 个特征分别有 5 个和 6 个可能的取值,那么其 VC 维为 9.一般情况下,$d_{\Theta}=1+\sum_{i=1}^{m}|Feature_i|$,其中,$|Feature_i|$是第 i 个特征的取值可能的数量.



(a) 特征 Title 无限取值可能 (b) 所有特征都有限 (c) 添加连接子句

Fig.3 Examples of VC dimension of MLN classification model

图 3 MLN 分类模型的 VC 维实例

另一种分类子句集的方式是将所有特征谓词的所有可能的情况都联合起来,与类别的所有情况形成一个合取的子句集,如:

*feature_i(row,+value_i)^feature_j(row,+value_i)^…^feature_n(row,+value_i)^*class(row,+value_k).

该子句集实际上是对每一个特征的每一种取值建立一个规则来拟合数据.假设关系联合表中有 m 个特征,那么在由这些特征组成的 $m\times m$ 的空间上,除非两个点的坐标相同,否则任意两个样本点都能通过调整相应规则的权重来区分,即规则集合的分类能力的最小粒度为一个样本点(一个规则能够确定一个样本点的类别情况,而不会影响其他点),因此,该子句集的 VC 维为无穷.

3.2.3 关联已确定的关系分类问题

集体分类的任务是在给定若干对象的属性以及它们之间的关系时,来判断这些对象所属的类别.作为关系模型的 MLN,能够很容易地合并相关的信息,并且能够一次性地推断出所有的对象类别.

关系数据与一般分类数据不同之处在于:其存在着直接特征和间接特征,并伴随着关系自相关特性.直接特征将一个对象的标签和另一个对象标签直接相连.如果没有附加的子句和谓词,则该连接关系无法用传统的 i.i.d 分类模型得出.而间接特征是通过属性的一串连接关系后将两个对象的标签连接起来的关系,可以通过连接相同的特征值的两个对象来实现.

定理 3 成立的一个假设为数据的连接是由数据所确定,而不受关系模型的影响.在马尔可夫逻辑网的分类模型中,这种假设可以由以下任意一个子句集实现:

class(row1,+value_c1)^LinkTo(row1,row2)⇒class(row2,+value_c2)

class(row1,+value_c)^LinkTo(row1,row2)⇒class(row2,+value_c)

class(row1,+value_c)^LinkTo(row1,row2)⇒class(row2,value_c)

其中,+value_c1,+value_c2 潜在地捕获连接模式,简单地包含 value_c,+value_c 子句仅仅有 5 个需要学习的权重,每一个针对一个类别标签.value_c,+value_c 子句仅仅表示同质性的依赖,也就是对象有相同的标签才会关联,但是允许非常不同的标签之间的关联.最后一个子句 value_c,value_c 是最简单的规则,对于所有样本仅仅学习一个共享的权重.

每一个子句都有能力去分类关联的样本.添加了连接子句到传统分类问题的子句中以后,将增加整体的 VC 维.但是由我们的观察得出,当添加了连接字句以后,传统分类问题的子句的 VC 维仅仅增加了 1.这主要是因为连接的数据限制了模型可能标记的样本数,如图 3(c)所示.更多的观察结果下:

- 当关系分类模型的 VC 维大于训练样本的数量时,该模型能够打散整个训练集的样本,这时不用添加额外的连接性子句就能获得很好的训练准确度;相反,如果添加了连接性子句,则有可能会造成训练集和测试集的误差增大.
- 当关系分类模型的 VC 维小于训练样本数量时,该模型不能打散所有的样本.这时如果能够找到未打散的样本和打散了的样本之间正确的关联,那么即使 VC 维并没有增加很多,但是由于模型更加拟合数据,因此还是有可能提高分类准确度的.

3.2.4 关联需要学习的关系分类

有时候,数据之间的关联并没有显示地给出,这时就需要关系分类模型学习得到.这个过程一般被称为结构学习,马尔可夫逻辑网连接直接特征可以通过如下的子句形式来实现:

feature_i(row1,+v)^feature_j(row2,+v)⇒class(row1,+value_k)⇔class(row2,+value_k)

class(row1,+value_k)⇔class(row2,+value_k)

class(row2,+value_k)⇔class(row3,+value_k)

class(row1,+value_k)⇔class(row3,+value_k)

其中,后 3 个子句为传递子句集,描述了对象类别的传递性.该子句集能够将样本点划分成多个块,但本身并没有将样本打散的能力,因此,该子句集的 VC 维为 0.虽然该子句集有能力连接相同特征值的两个对象,但是当对象的特征值都不相同时,无法连接两个对象,因此,该子句集的关系维数为 1.

另一种可能的连接方法是:将以上所有特征谓词的所有可能情况联合起来,推出两两对象的连接情况,如:

*feature_i(row,+value_i)^feature_j(row,+value_i)^…^feature_n(row,+value_i)^*class(row,+value_k)⇒

class(row1,+value_k)⇔ class(row2,+value_k)

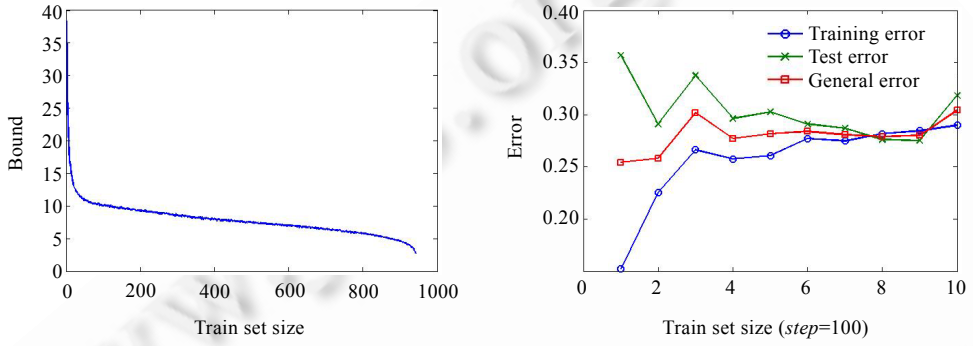
再加上传递子句集,这样的子句集合的 VC 维仍然是 0,但其关系维是∞.

以上两个例子为 MLN 模型的一些特殊情况.在实际的 MLN 模型学习中,由于子句长度和子句个数的限制,模型的 VC 维和关系维都可能是有限的.

3.2.5 实验结果

我们在 IMDB 数据集上做了一个简单的实验.该数据集[21]创建自 IMDB.com 的数据库,描述一个电影领域.其中,对象包括 movies,actors 和 directors.关系如 WorkedIn(person,movie),Actor(person),等等.我们挑选了 30%的数据作为测试集,剩余的 70%的数据作为训练集用上节提到的模型进行训练.该模型的 VC 维是有限的,因为在我们的设置下,所有的特征都是有限的.测试集有 245 个样本,训练集有 945 个样本.分类的具体任务是确定一个演员的性别基于指导他的导演.导演经常指导一个特定风格的电影,因此很可能需要特定性别的演员.

我们从训练集随机选择 N 个样本,为了获得精确的关于 k,d 的统计信息,每次采样将会重复 1 000 次.然后根据这些统计信息计算学习的界限,如图 4(a)所示.目前的结果显示:当样本数量为 945 时,界限仍然是无意义的,但是界限整体呈下降趋势,这说明模型是可学习的.图 4(b)显示:当模型具有有限 VC 维的时候,随着训练集数据量的增加,测试集的误差减少.这是由于模型的泛化能力增强了.这个实验结果与我们得出的学习界限一致.



(a) 界限 (b) 误差

Fig.4 Classification error and bound of MLN model on IMDB dataset

图 4 MLN 模型的界限和其实际分类误差

4 结论与展望

本文将传统机器学习的一般学习界限扩展到关系分类的学习问题中.首先引入关系自相关系数 d 以及独立子集个数 k 的统计信息函数,推导出当模型假设空间有限情况下的学习界限.然后,扩展传统机器学习的对称性引理到关系分类环境中,推导出当模型假设空间无限情况下的学习界限.接着,提出一个衡量关系模型关联数据能力的复杂性度量——关系维,证明了该复杂度和关系模型的生长函数之间的关系,推导出在有限 VC 维和有限关系维下的学习界限.本文还分析了该界限可学习和有意义的条件,并对界限的可行性进行了详细的分析.最后,分析了基于马尔可夫逻辑网的传统学习界限和关系分类中的学习情况.分析和实验结果表明,我们提出的界限能够解释实际关系分类中遇到的一些问题.

目前,大多数的关系学习模型研究集中在分析一个数据集作为整体输入模型的情况,还有一些研究考虑怎样切分数据为训练集合和测试集合.在这些研究中,训练集的大小是有限的,这意味着 k 和 d 的统计信息容易获得,那么我们提出的界限比较容易建立.然而,另外一些研究考虑输入的数据实际上本身也是来源于一个更大的网络,在这种情况下,N 是无限的,那么 k 和 d 的统计信息的稳定性难以确定,因此,所得到的模型也不具有很强的稳定性,界限也将难以建立.

下一步我们将研究该界限在更多的关系分类模型中的应用,以及用实验来获取关系模型 VC 维和关系维的近似值,并与文献[22]提出的关系分类学习的界限进行比较.如何在该界限的指导下建立一个稳定的、可学习的

关系分类模型,也是下一步的研究方向.

References:

[1] Jensen D, Neville J, Gallagher B. Why collective inference improves relational classification. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 593−598. [doi: 10.1145/1014052.1014125]

[2] Neville J, Jensen D. Collective classification with relational dependency networks. In: Proc. of the 2nd Int'l Workshop on Multi-Relational Data Mining. 2003. 77−91.

[3] Schmidt DC. Learning probabilistic relational models. In: Proc. of the Relational Data Mining. New York: Springer-Verlag, 2000. 307−333.

[4] Costa VS, Page D, Cussens J. CLP (BN): Constraint logic programming for probabilistic knowledge. In: Luc R, Paolo F, *et al*., eds. Proc. of the Probabilistic Inductive Logic Programming. Springer-Verlag, 2008. 156−188. [doi: 10.1007/978-3-540-78652-8_6]

[5] Kersting K, Raedt LD. Towards combining inductive logic programming with Bayesian networks. In: Proc. of the 11th Int'l Conf. on Inductive Logic Programming. Springer-Verlag, 2001. 118−131. [doi: 10.1007/3-540-44797-0_10]

[6] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006,62(1-2):107−136. [doi: 10.1007/s10994-006-5833-1]

[7] Getoor L. Tutorial on statistical relational learning. In: Proc. of the 15th Int'l Conf. on Inductive Logic Programming (ILP). Springer-Verlag, 2005. 415−415. [doi: 10.1007/11536314_26]

[8] Mihalkova L, Richardson M. Speeding up inference in statistical relational learning by clustering similar query literals. In: Proc. of the 19th Int'l Conf. on Inductive Logic Programming (ILP). Springer-Verlag, 2010. 110−122. [doi: 10.1007/978-3-642-13840-9_11]

[9] Singla P, Domingos P. Entity resolution with Markov logic. In: Proc. of the Int'l Conf. on Data Mining (ICDM). Hong Kong: Institute of Electrical and Electronics Engineers Inc., 2007. 572−582. [doi: 10.1109/ICDM.2006.65]

[10] Kontorovich LA, Ramanan K. Concentration inequalities for dependent random variables via the martingale method. The Annals of Probability, 2008,36(6):2126−2158. [doi: 10.1214/07-AOP384]

[11] Mohri M, Rostamizadeh A. Stability bounds for non-iid processes. Advances in Neural Information Processing Systems, 2007,20: 1025−1032.

[12] Ralaivola L, Szafranski M, Stempfel G. Chromatic PAC-Bayes bounds for non-iid data. In: Proc. of the 12th Int'l Conf. on Artificial Intelligence and Statistics. 2009. 416−423.

[13] Ralaivola L, Szafranski M, Stempfel G. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. Journal of Machine Learning Research, 2010,11(3):1927−1956.

[14] Janson S. Large deviations for sums of partly dependent random variables. Random Structures & Algorithms, 2004,24(3):234−248. [doi: 10.1002/rsa.20008]

[15] London B, Huang B, Getoor L. Improved generalization bounds for large-scale structured prediction. In: Proc. of the NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks. 2012.

[16] Dhurandhar A, Dobra A. Distribution-Free bounds for relational classification. Knowledge and Information Systems, 2012,31(1): 55−78. [doi: 10.1007/s10115-011-0406-4]

[17] Dhurandhar A. Auto-Correlation dependent bounds for relational data. In: Proc. of the 11th Workshop on Mining and Learning with Graphs. Chicago, 2013.

[18] Xu CF, Hao CL, Su BJ, Lou JJ. Research on Markov logic networks. Ruan Jian Xue Bao/Journal of Software, 2011,22(8): 1699−1713 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4053.htm [doi: 10.3724/SP.J.1001.2011.04053]

[19] Liu YB, Yang BR, Li GY, Liu YH. Joint inference open information extraction based on Markov logic networks. Computer Science, 2012,39(9):202−205 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2012.09.045]

[20] Silva VA, Polastro RB, Cozman FG. Learning classifiers with Markov logic network. In: Congresso Brasileiro de Automática (CBA 2008). 2008.

[21] Mihalkova L, Huynh T, Mooney RJ. Mapping and revising Markov logic networks for transfer learning. In: Proc. of the National Conf. on Artificial Intelligence (AAAI). Vancouver: American Association for Artificial Intelligence, 2007. 608−614.

[22] London B, Huang B, Taskar B, Getoor L. Collective stability in structured prediction: Generalization from one example. In: Proc. of the 30th Int'l Conf. on Machine Learning (ICML 2013). 2013.

附中文参考文献:

[18] 徐从富,郝春亮,苏保君,楼俊杰.马尔可夫逻辑网络研究.软件学报,2011,22(8):1699−1713. http://www.jos.org.cn/1000-9825/4053.htm [doi: 10.3724/SP.J.1001.2011.04053]

[19] 刘永彬,杨炳儒,李广源,刘英华.基于马尔可夫逻辑网的联合推理开放信息抽取.计算机科学,2012,39(9):202−205. [doi: 10.3969/j.issn.1002-137X.2012.09.045]

王星(1981−),男,重庆人,博士生,主要研究领域为网络与信息安全,网络舆情监控,知识迁移.

E-mail: yeahwx@gmail.com

方滨兴(1960−),男,博士,教授级高级工程师,博士生导师,CCF 高级会员,中国工程院院士,主要研究领域为计算机体系结构,信息安全,计算机网络.

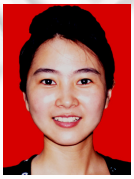
E-mail: fangbx@cae.cn

张宏莉(1973−),女,博士,教授,博士生导师,CCF 会员,主要研究领域为网络与信息安全,网络测量.

E-mail: zhanghongli@hit.edu.cn

何慧(1974−),女,博士,副教授,CCF 会员,主要研究领域为网络与信息安全.

E-mail: hehui@hit.edu.cn



赵蕾(1989−),女,硕士生,主要研究领域为网络舆情监控,知识迁移.

E-mail: zhaoleigogogo@163.com