科学通报 2009 年 第 54 卷 第 11 期: 1618 ~ 1625
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

基于混合正则化的无标签领域的归纳迁移学习

庄福振[①③]，罗平[②]，何清[①]，史忠植[①]

① 中国科学院计算技术研究所智能信息处理重点实验室，北京 100190;
② 惠普中国实验室，北京 100084;
③ 中国科学院研究生院，北京 100049
E-mail: zhuangfz@ics.ict.ac.cn

2008-07-08 收稿, 2008-11-22 接受
国家自然科学基金(批准号: 60435010, 60675010)、国家高技术研究发展计划(编号: 2006AA01Z128, 2007AA01Z132)、国家重点基础研究发展计划(编号: 2007CB311004)和国家科技支撑计划(编号: 2006BAC08B06)资助项目

摘要 近年来迁移学习已经引起了越来越广泛的兴趣，签数据以及源领域数据是不同分布的分类问题，且建立一个归纳分类模型对新来的目标数据进行预测. 首先分析了直推式迁移学习(transductive transfer learning)中存在的类别比例漂移问题，然后提出归一化的方法使得预测的类别比例接近于实际样本类别比例. 更进一步，提出了一种基于混合正则化框架的归纳迁移学习算法. 其中包括目标领域分布结构的流形正则化，预测概率的熵正则化，以及类别比例的期望正则化. 这个框架被用于从源领域到目标领域学习的归纳模型中. 最后，在实际文本数据集上的实验结果表明，提出的归纳迁移学习模型是有效的，同时该模型可以直接对新来的目标数据进行预测.

关键词
迁移学习
归纳学习
直推式学习
混合正则化

分类学习在智能信息处理中起着关键性的作用，其中包括Web网页，图像以及视频的处理等. 传统的分类学习研究假设标签数据与无标签数据来源于相同数据分布，但是在实际中无标签数据可能来自于不断变化的但语义相关的不同信息源，因此现有的分类方法不能很好处理这种情况. 在这种情况下，假设标签数据 $D_s=\{x_i^{(s)},y_i^{(s)}\}|_{i=1}^{n_s}$ 是来自于服从某种数据分布的源领域数据，而无标签数据 $D_t=\{x_i^{(t)}\}|_{i=1}^{n_t}$ 是采样于不同于源领域数据分布的目标领域. 源领域与目标领域之间这种数据分布不匹配的分类问题引起了研究人员对迁移学习广泛的兴趣与研究[1~13].

到目前为止，大多数迁移学习方面的工作都是处理目标领域数据里面有少量标签数据的问题，比如文献[1,2,5~7]. 在这种情况下，目标领域中的少量标签数据被用来改变源领域中标签数据的权重，或者使得在源领域数据中训练得到的模型适合于无标签的目标领域数据. 这方面的工作包括Migratory-Logit[5]，基于提升技术的迁移学习[1]，以及交叉领域适应SVM方法[6]. Wu等人[7]提出了一种新的方法，利用与测试数据不同分布的训练数据来提高SVM算法在测试数据集上的性能.

另外一方面，要标注目标领域里面的数据也是相当困难的，即使是少量的，因为这也要消耗很大的人力和时间，所以要求迁移学习方法能够处理完全没有标签的目标领域数据. 对于这个问题，Dai等人[3]提出了基于同步聚类(co-clustering)的分类方法，该方法通过对类别和特征进行同步聚类，实现类别标签的迁移. Xing等人[4]提出了一种桥接精化(bridged refinement)的迁移学习方法，该方法在精化的过程中不断地修正由源领域训练得到的模型在测试数据集上的预测类别，从而获得较高预测准确率. 很明显，基于桥接精化的迁移学习方法是一种直推式(transductive)的转换方法，它不能产生分类器，只能对模型精化中的目标领域数据进行预测，而不能对新来的样本进行直接判别. 对于新来的数据，最原始的方法就是重新进行桥接精化的过程，这对于整个

引用格式: 庄福振，罗平，何清，等. 基于混合正则化的无标签领域的归纳迁移学习. 科学通报, 2009, 54: 1618~1625
Zhuang F Z, Luo P, He Q, et al. Inductive transfer learning for unlabeled target-domain via hybrid regularization. Chinese Sci Bull, 2009, 54, doi: 10.1007/s11434-009-0171-x

学习过程来说，效率很低.

在本文中，我们首先分析了直推式迁移学习中的桥接精化方法，发现在迭代的过程中预测得到的样本比例在不断发生变化，而且这种类别漂移很大程度上影响了直推式迁移学习桥接精化算法的性能. 在一些实际应用中，对应测试样本的实际比例(各个类别的样本数除于总的样本数)是可以得到的，比如可以通过领域知识[1]或者基于统计估计得到. 一个实际的例子，我们可以统计从网上抓取下来的网页，新闻网页大约占 20%，因此可以在桥接精化算法中加入先验知识——测试样本的实际类别比例来提高算法的分类性能.

对于目标领域都是无标签数据的情况，我们提出了一种归纳迁移学习算法. 该算法与其他方法的不同在于，不仅能够处理完全无标签的目标领域数据，而且能够产生分类模型，对新来的测试样本进行直接预测. 它包括两个阶段，首先从源领域数据学习得到一个分类模型h_s，该模型代表的是从源领域数据中学到的知识；在第二阶段中，通过把无标签数据加入到模型h_s的精化中，从而得到最终的模型 h_t 能够很好地对目标领域数据进行预测. 我们提出的混合正则化框架包括三个正则化准则：流形正则化[14]、熵正则化[15]、期望正则化[16] (这些正则化准则将在第 3 部分进行详细介绍)，该框架利用在源领域数据中训练得到的初始模型h_s作为第二阶段优化的初始值，然后通过非线性数值优化技术使得该混合正则化框架收敛到一个局部最优点h_t. 实际文本分类问题的结果表明，本文提出的归纳迁移学习方法是有效的，且优于以往直推式的迁移学习方法.

1 直推式迁移学习

1.1 桥接精化方法

这一部分主要描述直推式迁移学习中的桥接精化方法，这个算法主要包括两个阶段迭代的桥接精化过程.

假设概率矩阵为 $T \in \mathrm{R}_{+}^{n\times|c|}$，其中 R_{+} 表示非负实数，$|c|$表示数据集的类别数，n 为样本的总数，那么 T_{ij} 表示为第 i 个样本属于第 j 类的概率，K 是近邻个数. M 为数据集的邻接矩阵，

$$M_{ij}=\begin{cases}0, & \text{如果 } x_j \text{不是} x_i \text{的} k\text{-近邻},\\ \dfrac{1}{K}, & \text{如果 } x_j \text{是} x_i \text{的} k\text{-近邻}.\end{cases} \tag{1}$$

两个样本 x_i 和 x_j 之间的相似度由余弦距离

$$\cos(x_i,x_j)=\frac{x_i^T x_j}{\|x_i\|\cdot\|x_j\|} \tag{2}$$

度量. 每次迭代中，样本都从近邻的样本中吸收一部分类别信息，同时也保留自身的部分类别信息，迭代公式如下:

$$T_{i\cdot}^{m+1}=\alpha\sum_{j:x_j\in N_i}\frac{T_{j\cdot}^{m}}{K}+(1-\alpha)T_{j\cdot}^{0}, \tag{3}$$

其中 $T_{i\cdot}^{m+1}$ 表示概率矩阵 T 经过 $(m+1)$ 次迭代后得到的第 i 行的值，N_i 是第 i 个样本所有 k -近邻的集合，$0<\alpha<1$ 是近邻类别信息与样本自身类别信息之间的平衡参数. 以上的迭代公式也可以写成下面矩阵计算形式:

$$T^{m+1}=\alpha MT^{m}+(1-\alpha)T^{0}. \tag{4}$$

可以证明得到概率矩阵收敛于下式:

$$T^{*}=(1-\alpha)(1-\alpha M)^{-1}T^{0}, \tag{5}$$

详细的理论分析证明见文献[17].

实际上，我们发现该桥接精化过程中每个阶段都需要两个输入参数，邻接矩阵 M 和概率矩阵 T^0. 在第一阶段的输入中，邻接矩阵 M 是在所有数据集上(包括源领域和目标领域数据)的邻接矩阵，T^0 是初始训练模型 h_s 在所有数据集上(包括源领域和目标领域数据)的预测概率矩阵. 在第二阶段，M 仅仅是在目标领域数据集上的邻接矩阵，T^0 是第一阶段迭代的结果，但只包括目标领域数据集部分. 因此，桥接精化方法的主要思想，首先考虑了所有数据 $D=D_s\cup D_t$ 上的流形结构，然后是目标领域数据 D_t 上的流形结构来提高预测结果的准确率.

1.2 强桥接精化方法

本文中，我们考查了概率矩阵在迭代精化过程中的性质，发现概率矩阵 T 的每一行的和保持不变，即 $\sum_{j=1}^{|c|}T_{ij}=1$. 但是 T 的每一列之和 $s=\sum_{i=1}^{n}T_{ij}$ 却在不断地变化，且 s 在一定程度上反映了预测样本属于类别 j 的样本数. 在实际中，待测样本中属于各个类别样本的数目是固定的，也就是我们期望 s 趋近于实际的样本数.

对于二类分类问题，

$$r^{p}=\frac{\sum_{i=1}^{n}T_{i1}}{\|T\|}, \tag{6}$$

$$r^{n}=\frac{\sum_{i=1}^{n}T_{i2}}{\|T\|}, \tag{7}$$

分别表示正负类样本在数据集中所占有的类别比例, 其中

$$\| T \| = \sum_{j=1}^{2} \sum_{i=1}^{n} T_{ij}, \quad (8)$$

n 是数据集的样本数. 如图 1 所示, 问题 10(表 1 给出了该问题的详细描述)的正样本类别比例 r^p($r^n = 1 - r^p$)随着迭代过程在不断变化. 我们可以看到 r^p 最终收敛到 0.571, 与实际的比例值 0.748 有很大的差异, 这可能导致算法的性能变差.

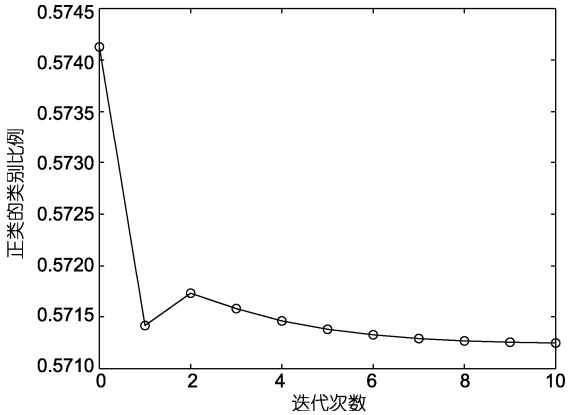

图 1 问题 10 的正样本类别比例在迭代过程中的变化

在Xing等人[4]的方法中, 把类别比例都归一化为 1:1, 这虽然可以处理类别比例平衡(正负样本类别比例基本相同)的情况, 但对于类别比例严重不平衡的情况是不适用的. 考虑在提供实际类别比例的情况下, 算法 1 (附录A)给出了加入类别比例归一化的桥接精化方法. 在每次迭代中都把迭代得到的类别比例归一化为实际的类别比例. 通过这个归一化技术, 可以把问题 10 的预测准确率由 80.47%提高到接近 92%, 大大提高了预测准确率.

2 基于混合正则化的归纳迁移学习

第 1 部分描述的直推式迁移学习方法不能够产生分类模型, 只能对精化过程的样本进行预测, 而不能对新来的测试样本直接预测. 但是在很多实际应用中, 都希望得到一个最终分类器对新来的样本进行直接预测, 因此我们提出了基于混合正则化的两阶段归纳迁移学习方法. 该方法可以推广到多类情况, 以下仅考虑两类问题. 由于我们的算法是基于逻辑回归[19]实现的, 下面我们将简要描述该分类模型.

2.1 逻辑回归

逻辑回归[19]是一种分类学习方法, 条件概率 $P(Y \mid X)$ 是在给定样本 X 的情况下, 求样本 X 属于类别 Y 的概率, 其中 Y 是离散型的值, 而 X 是任意的包含离散或者连续型随机变量的向量. 逻辑回归通过优化目标函数, 在训练数据集上估计参数模型. 当 Y 是布尔型时, 分类模型如下:

$$P(y = \pm 1 \mid x; w) = \sigma(y w^{\mathrm{T}} x) = \frac{1}{1 + \exp(-y w^{\mathrm{T}} x)}, \quad (9)$$

其中 w 是参数模型. 在最大后验估计原则下, 参数模型 w 通过拉普拉斯先验估计. 给定训练数据集 $D = \{x_i, y_i\}|_{i=1}^{N}$, 可以通过优化最大化(10)式来求参数模型 w,

$$\sum_{i=1}^{N} \log \frac{1}{1 + \exp(-y_i w^{\mathrm{T}} x_i)} - \frac{\lambda}{2} w^{\mathrm{T}} w, \quad (10)$$

由于该函数对变量 w 是一个凹函数, 因此可以通过非线性数值优化方法求得全局最优解. 当得到参数模型以后, (9)式就可以用来计算待测样本分别属于正负类的概率.

2.2 正则化准则

本文提出的混合框架包括 3 种正则化准则: 流形正则化[14], 熵正则化[15]以及期望正则化[16]. 在以往的研究中, 这些准则被用于半监督学习中, 开发和利用标签和无标签数据的内部本质结构. 但在本文中, 我们把这些准则用于迁移学习中, 然后作用于与源领域具有不同数据分布的目标领域数据 $D_t = \{x_i^{(t)}\}|_{i=1}^{n_t}$ 上(其中 n_t 是目标领域数据集的样本数). 假设参数模型为 w, 那么(9)式的函数 σ 同样被用来表示条件概率 $P(y = 1 \mid x)$.

流形正则化 Belkin 等在半监督学习的研究中提出了有效开发和利用标签和无标签数据的流形结构的方法, 该准则要求样本的类别标签与其周围样本的类别标签是相似的. 我们把该准则运用到完全无标签的目标领域迁移学习中, 可以通过最小化以下式子来实现:

$$g_m(w) = \frac{1}{n_t} \sum_{i=1}^{n_t} \left[\frac{1}{K} \sum_{k=1}^{K} \sigma(w^T x_{i_k}) - \sigma(w^T x_i) \right]^2, \quad (11)$$

其中 K 是样本 x_i 的近邻个数, x_{i_k} 是样本 x_i 的第 k 个近邻($1 \le k \le K$). 我们可以采用任何一种相似度度量方式来计算样本之间的相邻关系.

熵正则化 Grandvalet 等提出的熵正则化实现了对样本 x_i 的预测概率向量 $p_i = (p_{i1}, \cdots, p_{i|c|})$ 的熵最小化, 其中 p_{ij} 是样本 x_i 属于类别 c_j 的概率, $|c|$ 是样

本的类别数. 熵正则化准则主要是基于每个样本都属于且仅属于一种类别的事实，我们都希望最终得到真实的概率预测向量. 对于二类问题，熵正则化等价于最小化以下式子:

$$g_c(w)=-\frac{1}{n_t}\sum_{i=1}^{n_t}\left[\sigma(w^{\mathrm{T}}x_i)-\frac{1}{2}\right]^2. \tag{12}$$

期望正则化　Mann 等提出的期望正则化准则可以使预测得到的结果逼近于一些先验知识，比如样本的实际类别. 也就是说，可以使得预测得到的样本类别比例接近于实际的类别比例. 形式化表示如下:

$$g_e(w)=\frac{1}{n_t}\left[\sum_{i=1}^{n_t}\sigma(w^{\mathrm{T}}x_i)-r\cdot n_t\right]^2, \tag{13}$$

其中 r 是正样本的实际比例.

2.3　两个阶段的归纳迁移学习

我们提出的归纳迁移学习方法包括两个阶段，首先是训练初始的分类器模型，然后是对初始分类模型的精化.

第一阶段　训练初始分类器模型 h_s. 假设源领域的标签数据为 $D_s=\{x_i^{(s)},y_i^{(s)}\}|_{i=1}^{n_s}$，然后用监督学习算法逻辑回归[19]来学习得到在源数据集上的初始模型 h_s.

第二阶段　通过混合正则化框架精化初始分类器模型 h_s. 给出目标领域数据 D_t，我们优化以下目标函数 f:

$$f(w)=w^{\mathrm{T}}w+\alpha\cdot g_m+\beta\cdot g_c+\gamma\cdot g_e, \tag{14}$$

其中 α，β 和 γ 为这些正则化准则之间的平衡因子，g_m，g_c 和 g_e 分别为(11)~(13)式的定义.

本文提出的混合正则化框架，与半监督学习的不同在于，我们没有考虑源领域标签数据上的 log 似然，如(10)式中的第一项. 这是因为在半监督学习中，标签数据和无标签数据都是来自于同种分布，自然地可以共享相同的分类模型. 但是在迁移学习中，源领域的数据和目标领域的数据具有不同的数据分布，因此不可能找到一个优化后的分类模型同时在两种数据集上都表现很好. 所以，我们首先在源领域的标签数据上训练初始模型，然后只在目标领域的无标签数据上精化该模型，最后得到优化后的分类模型可以在目标领域上表现得很好且能直接预测新来的样本.

为了解决该优化问题，我们给出函数 g_m，g_c 以及 g_e 对变量 w 的偏导数，

$$\nabla_w g_m=\frac{2}{n_t}\cdot\sum_{i=1}^{n_t}\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(w^{\mathrm{T}}x_{i_k})-\sigma(w^{\mathrm{T}}x_i)\right)$$
$$\times\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(w^{\mathrm{T}}x_{i_k})(1-\sigma(w^{\mathrm{T}}x_{i_k}))x_{i_k}-\sigma(w^{\mathrm{T}}x_i)(1-\sigma(w^{\mathrm{T}}x_i))x_i\right), \tag{15}$$

$$\nabla_w g_c=-\frac{2}{n_t}\cdot\sum_{i=1}^{n_t}\left(\sigma(w^{\mathrm{T}}x_i)-\frac{1}{2}\right)\sigma(w^{\mathrm{T}}x_i)(1-\sigma(w^{\mathrm{T}}x_i))x_i, \tag{16}$$

$$\nabla_w g_e=\frac{2}{n_t}\left(\sum_{i=1}^{n_t}\sigma(w^{\mathrm{T}}x_i)-r\cdot n_t\right)\left(\sum_{i=1}^{n_t}\sigma(w^{\mathrm{T}}x_i)(1-\sigma(w^{\mathrm{T}}x_i))x_i\right), \tag{17}$$

因此目标函数 f 的偏导数为

$$\nabla_w f=2\cdot w+\alpha\cdot\nabla_w g_m+\beta\cdot\nabla_w g_c+\gamma\cdot\nabla_w g_e. \tag{18}$$

由于目标函数 f 既不是凸函数也不是凹函数，因此很难求解得到最优值. 但是在给定初始值 h_s 的情况下，可以很容易用非线性数值优化方法得到局部最优解，且这个解是优于初始值的. 在本文中，我们采用了共轭梯度方法来求解这个目标函数，详细的共轭梯度优化求解过程见算法 2 (附录 B). 另外，我们采用 Matlab 本身自带的函数 ***fminunc*** (求无约束最小值点函数)来求解算法 2 中步骤 3 的单变量优化问题.

3　实验结果

3.1　实验数据

我们重新构造数据集 20 *newsgroups*[1)]使其符合本文中迁移学习问题的要求，该数据集有两层的层次结构. 假设 A 和 B 分别表示数据集中顶层的两个类别，A_1, A_2 和 B_1, B_2 分别是属于类别 A 和 B 的第二层的子类别. 我们构造源领域和目标领域数据集如下，让 $A.A_1$ 和 $B.B_1$ 分别为源领域数据集中的正负类样本；而 $A.A_2$ 和 $B.B_2$ 分别为目标领域数据集中的正负类样本. 我们得到 12 个分类问题，如表(附录 C)所述. 同时表中也列出了正样本在目标领域数据集中的比例 r^p. 每个文档都用 $tf\cdot idf$ 方法表示成一个向量，文档频率的阈值设置为 5 来选择特征.

3.2　性能比较

在本文中，与基于混合正则化框架的归纳迁移

1) http://people.csail.mit.edu/jrennie/20Newsgroups/

学习方法(IHR)比较的基线算法包括:

(ⅰ) 传统的分类学习算法: SVM[18]和逻辑回归(LR)[19].

(ⅱ) 直推式迁移学习方法: 桥接精化方法[4](BR)和加入类别比例的强桥接精化方法(PBR). 根据初始预测概率矩阵的不同, 桥接精化方法又可以分为 BR^{LR}, BR^{SVM}, PBR^{LR}和PBR^{SVM}. BR^{LR}和BR^{SVM}分别表示初始预测概率矩阵由逻辑回归和SVM算法预测得到的桥接精化方法, PBR^{LR}和PBR^{SVM}类似. 这些直推式方法中, 近邻的个数都设置为 $K=70$; SVM采用的是线性核, 其他的参数都采用默认值 1).

(ⅲ) 归纳迁移学习方法: 基于同步聚类的分类算法CoCC[3].

(ⅳ) 半监督学习方法 2): TSVM[20]和SGT[21].

算法 IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} 和 PBR^{SVM}: 我们比较了以上 7 个算法在 12 个分类问题上的性能, 每一种算法在 12 个分类问题上的准确率如图 2 所示, 以及每一种算法在 12 分类问题上的平均性能如表 1 所示. 为了验证算法性能的优越性, 我们还做了统计 t 测试(置信度为 95%), 从而发现: 1) IHR 算法在统计意义上, 相对于算法 LR, SVM, BR^{LR}, 以及 BR^{SVM} 有很大的提高; 2) IHR 方法相对于算法 PBR^{LR} 和 PBR^{SVM} 的优越性并不是很明显. 但是如表 2 所示, 从平均性能上看, IHR 优于算法 PBR^{LR} 和 PBR^{SVM}.

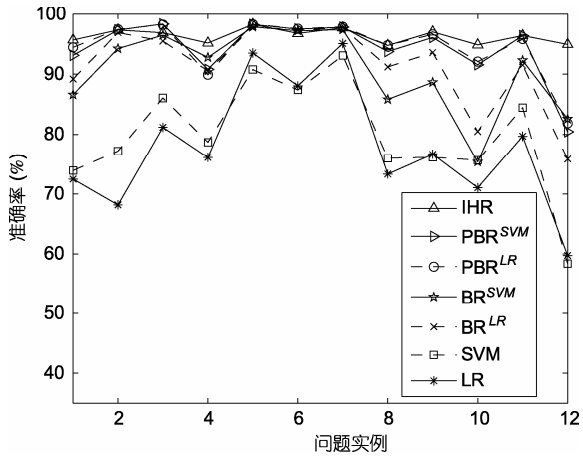

图 2 算法 IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} 和 PBR^{SVM} 在 12 个分类问题上的性能(%)比较($\alpha=0.4$, $\beta=15$, $\gamma=0.12$)

表 1 算法 IHR, LR, SVM, BR^{LR}, PBR^{LR}, BR^{SVM} 和 PBR^{SVM} 在 12 个分类问题上的平均性能(%) ($\alpha=0.4$, $\beta=15$, $\gamma=0.12$)

$\overline{\text{LR}}$	$\overline{\text{SVM}}$	$\overline{BR^{LR}}$	$\overline{BR^{SVM}}$	$\overline{PBR^{LR}}$	$\overline{PBR^{SVM}}$	$\overline{\text{IHR}}$
77.92	79.81	91.46	90.59	94.57	94.27	96.31

算法IHR, CoCC[3], TSVM[20]和SGT[21]. 为了体现出算法之间的可比性和公平性, 我们采用了Dai等[3]论文中的数据(数据的详细描述见其文章中的表 1), 结果如表 2 所示. 从表 2 可以看到算法IHR在所有数据集上的准确率都高于CoCC, TSVM和SGT, 再一次验证了IHR算法的优越性和有效性.

表 2 算法 TSVM, SGT, CoCC 和 IHR 之间的性能(%)比较($\alpha=0.4$, $\beta=15$, $\gamma=0.12$)

Data set	TSVM	SGT	CoCC	IHR
res *vs*. talk	96	90.9	96.4	98.71
rec *vs*. sci	93.8	93.8	94.5	97.15
comp *vs*. talk	90.3	97.2	98	98.28
comp *vs*. sci	81.7	72.1	87	96.65
comp *vs*. rec	90.2	95.3	95.8	97.04
sci *vs*. talk	89.2	91.7	94.6	96.15

3.3 参数影响

本文还考查了平衡参数 α, β 以及 γ 的不同设置对算法 IHR 的影响, 我们选取了 6 个分类问题作为实验数据. 在这些参数实验中, 我们记录了算法 IHR 的性能随其中任意一个参数变化的影响(另外两个参数取固定值). 结果如图 3 所示. 通过分析, 我们发现:

从图 3(a), 可以看到算法 IHR 的性能对参数 $\alpha\in[0,1]$ 是不敏感的, 从实验结果也可以看到流形正则化对某些分类问题是没有作用的, 如问题 5 和 6. 我们分析了问题 5 和 6 中出现的现象, 发现只要初始模型的预测结果已经有较好的流形结构性, 即样本自身的类别已经和近邻样本有很高的一致性, 这样流形正则化的作用就很小, 甚至不起作用. 因此, 我们认为影响算法 IHR 性能的不仅是参数的调节, 而且跟数据本身的特点也有很大关系.

从图 3(b)和图 3(c)可以看到参数 β 和 γ 对算法

1) 我们也对分类算法 SVM 和逻辑回归算法的参数进行了调节, 发现这并不能进一步提高算法 BR 和 PBR 的性能. 文章中采用的默认参数是比较好的, 总是可以获得较好的分类性能. 这里没有列出详细调参数的实验结果

2) TSVM和SGT的参数设置与Dai[3]等论文中的设置一致

IHR 有很大的影响，不过图 3(c)也表明当 γ 大于 0.05 的时候，算法的性能是稳定的.

为了证明算法 IHR 对参数设置的健壮性，我们放松参数 α，β 和 γ 的取值范围，而不是取以上实验中固定的参数值. 经过初步实验，设置参数的取值范围为 $\alpha \in (0,1)$，$\beta \in (0,30)$ 以及 $\gamma \in (0,0.5)$，然后评价算法在 12 个分类问题(表 1 描述)上的性能. 我们对参数随机采样 m (这里 $m=15$)种组合，并在每种参数组合下对所有的 12 分类问题平均算法准确率，结果如表 3 所示. 我们发现该平均性能与第 3.2 节($\alpha=$ 0.4, $\beta=$ 15, $\gamma=$ 0.12)给出的结果几乎是一样的；而且我们也看到在所有的参数采样情况下，算法 IHR 的性能都表现得很好，这也再一次验证了算法 IHR 的有效性和健壮性.

3.4 归纳式学习算法

本文提出的算法 IHR 与直推式算法的不同还在于该算法是归纳式的，可以产生最终分类器对新来的样本进行预测. 为了验证算法 IHR 在新来的数据集上的性能，我们依旧采用表中描述的 12 个分类问题. 对于每个分类问题中的目标领域无标签数据，我们随机采样(无放回采样)比例为 p 的数据构成新的数据集 D_t^1，剩下的构成数据集 D_t^2. 数据 D_t^1 用于对初始模型 h_s 的精化过程，而数据 D_t^2 用于测试精化得到的模型 h_t 的泛化能力. 我们记录了算法 IHR 在数据集 D_t^1 和 D_t^2 上准确率，同时也记录了不同采样比例 p 下算法 IHR 的分类性能，所有的结果如表 4 和图 4 所示.

从这些结果，我们发现：算法 IHR 在数据集 D_t^1 和 D_t^2 上的分类准确率几乎是一致的；用于精化过程的无标签数据集 D_t^1 越大，算法的泛化能力就越好. 当采样比例 $p \geq 0.6$ 时，算法在数据集 D_t^2 上的泛化能力已经达到 90%以上. 该结果表明，算法 IHR 在采样比例大于 60%时，对新来的测试数据就能表现出很好的泛化能力.

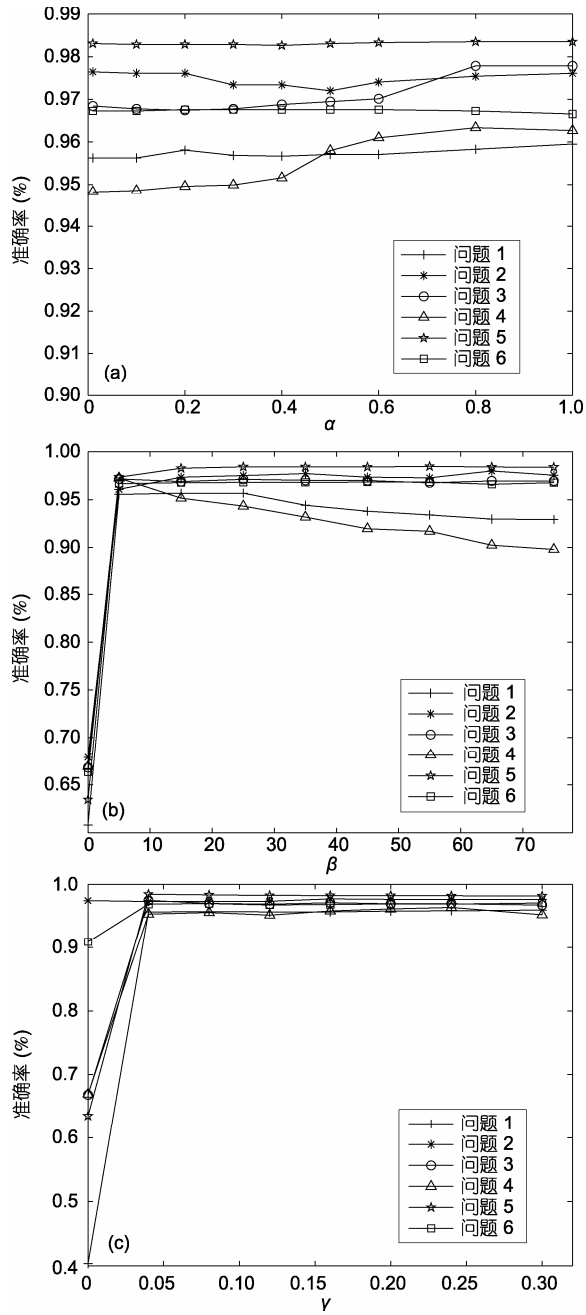

图 3 参数对算法 IHR 的影响

(a) α vs. IHR 的准确率($\beta=$ 15, $\gamma=$ 0.12)关系曲线图; (b) β vs. IHR 的准确率($\alpha=$ 0.4, $\gamma=$ 0.12)关系曲线图; (c) γ vs. IHR 的准确率($\alpha=$ 0.4, $\beta=$ 15)关系曲线图

4 相关工作

迁移学习处理的是训练样本与测试样本来自于不同数据分布的分类问题，通常，可以把以前所做的工作分为两类. 第一类，目标领域中有少量的标签数据，例如Liao等人[5]提出了一种方法，估计源领域中的每个样本与目标领域中少量标签数据之间的不匹配程度，并把该信息应用到逻辑回归中；Dai等人[1]扩展Boosting学习算法到迁移学习中，改变样本每次被采样的权重，即在迭代中源领域中的样本权重被减弱，而有利于模型训练的目标领域中的样本权重被加强. 他们还用 PAC 理论分析证明了该算法的有效性.

表 3 参数设置对算法 IHR 性能(%)的影响

试验编号	α	β	γ	问题编号												平均值
				1	2	3	4	5	6	7	8	9	10	11	12	
1	0.550	8.897	0.254	95.8	97.0	97.8	97.6	97.8	97.1	97.3	95.2	96.5	93.9	95.5	95.3	96.4
2	0.014	17.411	0.432	95.7	97.4	96.8	94.8	98.2	96.7	97.6	94.1	97.1	95.4	96.4	94.5	96.2
3	0.088	8.604	0.177	95.6	96.8	97.6	97.4	97.7	97.0	97.6	94.7	96.4	94.6	95.4	95.0	96.3
4	0.745	20.812	0.310	96.0	97.5	97.1	95.3	98.3	97.0	97.7	93.9	97.2	95.7	96.6	94.1	96.4
5	0.626	17.952	0.128	95.8	97.3	97.0	94.9	98.4	96.6	97.5	94.9	96.7	95.0	96.4	94.3	96.2
6	0.008	16.624	0.346	95.8	97.5	96.7	94.9	98.2	96.8	97.6	94.2	97.0	95.3	96.4	94.2	96.2
7	0.732	27.048	0.146	95.3	97.6	97.3	93.9	98.4	97.0	97.6	94.1	97.2	95.2	96.6	93.3	96.1
8	0.416	24.654	0.359	95.5	97.5	97.2	94.6	98.4	97.0	97.8	94.2	97.3	94.7	96.5	93.5	96.2
9	0.128	1.819	0.234	93.0	92.2	93.8	94.1	95.1	94.9	95.6	87.7	92.2	86.7	87.9	90.5	92.0
10	0.182	9.711	0.142	95.6	96.9	97.6	96.1	97.9	96.8	97.5	95.4	96.3	92.6	95.8	94.9	96.1
11	0.842	22.251	0.257	95.9	97.5	97.1	95.3	98.4	97.0	97.9	94.6	97.3	95.5	96.8	94.1	96.4
12	0.196	25.486	0.082	95.3	97.4	97.2	93.8	98.5	96.8	97.6	93.8	96.6	95.4	96.3	93.2	96.0
13	0.981	13.580	0.427	95.9	97.5	97.8	97.0	98.1	97.2	97.7	96.3	97.2	95.2	96.2	95.2	96.8
14	0.303	29.857	0.405	94.8	97.8	97.0	93.7	98.4	96.9	97.7	93.8	97.0	92.6	95.8	92.9	95.7
15	0.793	6.347	0.295	95.8	96.6	97.5	98.0	97.5	96.9	97.1	95.7	96.2	93.5	95.0	96.2	96.3
	0.4	15	12	95.7	97.3	96.9	95.2	98.3	96.8	97.7	94.7	97.0	94.9	96.4	94.9	96.3

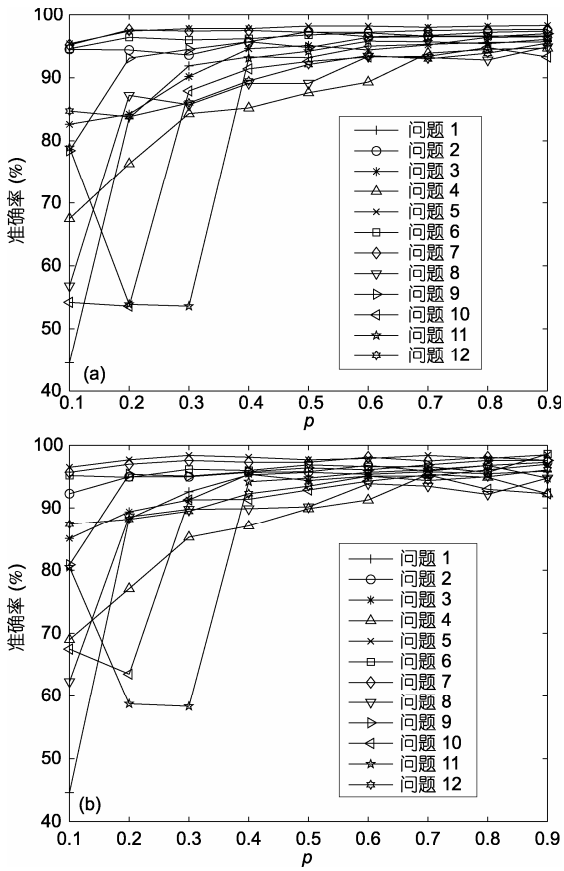

图 4 算法 IHR 在数据集 D_t^1, D_t^2 上的性能

(α = 0.4, β = 15, γ = 0.12)

(a) 采样比例 p vs. D_t^1 上的准确率关系曲线图; (b) 采样比例 p vs. D_t^2 上的准确率关系曲线图

在第二类迁移学习方法中，目标领域中的数据完全是没有标签的. 对于这种问题，Ben-David等人[9]分析了领域数据的表示，并提出了一个很好的模型，该模型不仅最小化分类模型在训练数据上的泛化误差，而且最小化源领域与目标领域之间的不同性. Ling等人[12]提出了一种新的光谱分类算法，该算法通过优化一个目标函数来寻找源领域中的监督信息与目标领域的本质结构之间的最大一致性. Mahmud等[13]从算法信息论的角度来研究迁移学习，该方法度量了不同任务之间的相关性，然后决定多少信息可以作迁移以及怎么迁移这些信息. Xing等人[4]提出了一种直推式迁移学习方法，该方法首先开发了所有数据集(包括源领域数据和目标领域数据)上的几何分布结构，然后再利用目标领域上的流形结构. 但是该方法不能产生分类器对新来的测试数据进行预测.

5 结论

本文提出了一种归纳迁移学习的方法，解决目标领域中都是无标签数据的情况. 首先我们分析了直推式迁移学习方法—桥接精化算法，发现样本类别比例在精化过程中的漂移，以及该算法在处理不平衡数据时的局限性. 然后在给出类别先验的情况下，我们提出了归一化的方法来解决该问题. 第二，提出了一种基于混合正则化的归纳迁移学习方法，包括流形正则化、熵正则化以及期望正则化等三种正则化准则. 与直推式迁移学习方法比较，该算法框架

可以得到最终分类器对新来的样本进行直接预测，另外，在实际数据集中的分类结果也验证了我们提出的算法的有效性．在以后的研究工作中，我们将探讨这些正则化准则的作用机理．

参考文献

1 Dai W Y, Yang Q, Xue G R, et al. Boosting for Transfer Learning. In: Ghahramani Z B, ed. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20—24, Corvalis, Oregon. ACM, 2007. 193—200

2 Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data. In: Ghahramani Z B, ed. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20—24, Corvalis, Oregon. ACM, 2007. 759—766

3 Dai W Y, Xue G R, Yang Q, et al. Co-clustering based classification for out-of-domain documents. In: Berkhin P, Caruana R, Wu X D, et al, eds. Proceeding of 13th ACM International Conference on Knowledge Discovery and Data Mining, 2007, Aug 12—15, San Jose, California. ACM, 2007. 210—219

4 Xing D K, Dai W Y, Xue G R, et al. Bridged refinement for transfer learning. In: Joost N K, Jacek K, Ramon L, et al, eds. Proceeding of 11th European Conference on Practice of Knowledge Discovery in Databases, 2007 Sep 17—21, Warsaw Poland. Springer, 2007. 324—335

5 Liao X J, Xue Y, Carin L, et al. Logistic regression with an auxiliary data source. In: Raedt L D, Wrobel S, eds. Proceeding of 22th International Conference on Machine Learning, 2007 Aug 7—11, Bonn, Germany. ACM, 2005. 505—512

6 Yang J, Yan R, Hauptmann A G, et al. Cross-domain video concept detection using adaptive SVMs. In Rainer L, Anand R P, Alan H, et al, eds. Proceeding of 15th International Conference on Multimedia, 2007 Sep 24—29, Augsburg Germany. ACM, 2007. 188—197

7 Wu P C, Dietterich T G. Improving SVM Accuracy by Training on Auxiliary Data Sources. In: Brodley C E, eds. Proceeding of 21th International Conference on Machine Learning, 2004 Jul 4—8, Banff, Alberta, Canada. ACM, 2004. 871—878

8 Mahmud M M H, Ray S, et al. Transfer learning using kolmogorov complexity: Basic theory and empirical evaluations. Technical Report, UIUC-DCS-R-2007-2875, Department of Computer Science, University of Illinois at Urbana-Champaign. 2007

9 Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation. In: Koller D, Singer Y, Platt J, et al, eds. Proceeding of Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA, 2007, (20): 137—144

10 Dai W Y, Xue G R, Yang Q, et al. Transferring naive bayes classifiers for text classification. In: John C, Peyman F, Simon P, et al, eds. Proceeding of 22nd Conference on Artificial Intelligence, 2007 Jul 22—26, Vancouver, British Columbia. AAAI Press, 2007. 540—545

11 Samarth S, Sylvian R. Cross domain knowledge transfer using structured representations. In: Proceeding of 21nd Conference on Artificial Intelligence, 2006 Jul 16—22, Boston, Massachusetts. AAAI Press, 2006

12 Ling X, Dai W Y, Xue G R, et al. Spectral domain-transfer learning. In: Li Y, Liu B, Sunita S, et al, eds. Proceeding of 14th ACM International Conference on Knowledge Discovery and Data Mining, 2008, Aug 24—27, Las Vegas, Nevada. ACM, 2008. 488—496

13 Mahmud M M H. On Universal Transfer Learning. In: Rocco M H, Servedio R A, Takimoto E, et al, eds. Proceeding of 18th International Conference on Algorithmic Learning Theory, 2007, Oct 1-4, Sendai, Japan. LNCS, 2007. 135—149

14 Belkin M, Niyogi P, Sindhwani V, et al. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res, 2006, 7: 2399—2434

15 Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceeding of 19th Conference on Neural Information Processing Systems, 2005 Dec 5—8, Vancouver, British Columbia. MIT Press, 2005. 529—536

16 Mann G S, McCallum A. Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization. In: Ghahramani Z B, eds. Proceeding of 24th International Conference on Machine Learning, 2007 Jun 20—24, Corvalis, Oregon. ACM, 2007. 593—600

17 Wang F, Zhang C S. Label propagation through linear neighborhoods. IEEE T Know Data Engin, 2008(20): 55—67

18 Joachims T. Making Large-scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A, et al, eds. Proceeding of Advances in Kernel Methods, 1999. MIT Press, Cambridge, 1999. 169—184

19 Davie H, Stanley L. Applied Logistic Regression. New York: Wiley, 2000

20 Joachims T. Transductive Inference for Text Classification Using Support Vector Machines. In: Bratko I, Dzeroski S, eds. Proceeding of 16th International Conference on Machine Learning, 1999 Jun 27—30, Bled, Slovenia. ACM, 1999. 200—209

21 Joachims T. Transductive Learning via Spectral Graph Partitioning. In: Tom F, Nina M, eds. Proceeding of 20th International Conference on Machine Learning, 2003 Aug 21—24, Washington DC. ACM, 2003. 290—297

附录 A

算法 **1:** 集成归一化技术的桥接精化方法

输入: 数据集 D, 初始未精化的概率矩阵 T^0, 平衡参数 α, 近邻个数 K, 以及所有样本类别的比例先验 $r=(r_1,\cdots,r_{|c|})$, 其中 $\sum_{i=1}^{|c|} r_i=1$.

输出: 精化后的概率矩阵 T.

步骤 **1:** 对每两个样本 x_i 和 x_j, 计算它们之间的相似度,

$$S_{ij}=\cos(x_i,x_j)=\frac{x_i^T x_j}{\|x_i\|\cdot\|x_j\|};$$

步骤 **2:** 对每个样本 x_i, 求出其 k 近邻 N_i;

步骤 **3:** 求出每个样本的 k 近邻后, 构造相邻矩阵 M.

步骤 **4:** 对每个类别 c_j 的预测概率向量做归一化,

$$T_{\cdot j}=T_{\cdot j}/(\|T_{\cdot j}\|/(r_j\cdot n));$$

步骤 **5:** 对每个样本 x_i, 计算第 $t+1$ 次迭代的概率,

$$T_{i\cdot}^{t+1}=\alpha\sum_{j:x_j\in N_i}M_{ij}T_{j\cdot}^{t}+(1-\alpha)T_{i\cdot}^{0};$$

对每个类别 c_j 的预测概率向量做归一化,

$$T_{\cdot j}^{t+1}=T_{\cdot j}^{t+1}/(\|T_{\cdot j}^{t+1}\|/(r_j\cdot n));$$

步骤 **6:** 循环步骤 5, 直到预测概率矩阵 T 收敛.

步骤 **7:** 返回 T.

其中 n 是样本的个数, r_j 是类别 c_j 在样本集 D 中的比例, $\|T_{\cdot j}\|=\sum_{i=1}^{n}T_{ij}$. 算法的收敛条件是 $|(\|T^{m+1}\|-\|T^{m+1}\|)|<\varepsilon$, $\varepsilon>0$.

附录 B

算法 **2:** 共轭梯度法求解本文提出的方法

输入: 初始分类器 $h_s=w_0$, 源领域的数据 $D_s=\{x_i^{(s)},y_i^{(s)}\}_{i=1}^{n_s}$, 近邻个数 K, 正样本的类别比例 r, 参数 α,β,γ 的值以及误差阈值 ε.

输出: 精化后的分类器 w.

步骤 **1:** 计算偏导数 $\nabla_w f(w_0)$, $\nabla_w f(w_0)=2w_0+\alpha\cdot\nabla_w g_m(w_0)+\beta\cdot\nabla_w g_c(w_0)+\gamma\cdot\nabla_w g_e(w_0)$;

如果 $\|\nabla_w f(w_0)\|<\varepsilon$, 则转步骤 7, 否则计算初始的搜索方向 d_0, $d_0=-\nabla_w f(w_0)$.

步骤 **2:** $k=0$;

步骤 **3:** 最小化以下式子, 求第 k 次迭代的最佳步长 λ_k,

$$f(w_k+\lambda_k d_k)=\min_{\lambda} f(w_k+\lambda d_k);$$

步骤 **4:** 得到第 k 次迭代的最佳步长 λ_k 后, 计算第 $k+1$ 次的分类器 $w_{k+1}=w_k+\lambda_k d_k$.

步骤 **5:** 计算偏导数 $\nabla_w f(w_{k+1})$, 如果 $\|\nabla_w f(w_{k+1})\|<\varepsilon$, 则转步骤 7, 否则计算第 $k+1$ 次的搜索方向 d_{k+1},

$$d_{k+1}=-\nabla_w f(w_{k+1})+\mu_k d_k;$$

其中 μ_k 由 Palak-Ribiere-Polyak 公式计算得到,

$$\mu_k=\frac{\nabla_w f(w_{k+1})^T[\nabla_w f(w_{k+1})-\nabla_w f(w_k)]}{\nabla_w f(w_k)^T\nabla_w f(w_k)}.$$

步骤 **6:** $k=k+1$, 转步骤 3.

步骤 **7:** 返回精化后的分类器 w, $h_t=w$.

附录 C

表 算法性能评价中的数据描述

数据集	源领域数据 D_s	目标领域数据 D_t	类别比例 r^p
问题 1	comp.graphics, sci.electronics, comp.os.ms-windows.misc, sci.crypt	comp.sys.mac.hardware, sci.med, comp.sys.ibm.pc.hardware, comp.windows.x, sci.space	0.60
问题 2	rec.autos, talk.politics.guns, rec.motorcycles, talk.politics.misc	rec.sport.baseball, rec.sport.hockey, talk.politics.mideast	0.68
问题 3	rec.autos, sci.med, rec.sport.baseball, sci.space	rec.motorcycles, rec.sport.hockey, sci.crypt	0.67
问题 4	rec.autos, sci.med, rec.sport.baseball, sci.space	rec.motorcycles, rec.sport.hockey, sci.electronics	0.67
问题 5	talk.religion.misc, talk.politics.mideast, comp.sys.mac.hardware, comp.graphics	comp.windows.x, comp.sys.ibm.pc.hardware, talk.politics.guns, talk.politics.misc, comp.os.ms-windows.misc	0.63
问题 6	comp.graphics, rec.sport.hockey, comp.sys.ibm.pc.hardware, rec.motorcycles	comp.windows.x, rec.autos, comp.os.ms-windows.misc	0.66
问题 7	comp.graphics, rec.sport.hockey, comp.sys.ibm.pc.hardware, rec.motorcycles	comp.windows.x, rec.sport.baseball, comp.os.ms-windows.misc	0.66
问题 8	sci.electronics, talk.politics.misc, sci.med, talk.religion.misc	sci.crypt, talk.politics.guns, sci.space	0.68
问题 9	sci.electronics, talk.politics.misc, sci.med, talk.religion.misc	sci.crypt, talk.politics.mideast, sci.space	0.68
问题 10	comp.os.ms-windows.misc, sci.crypt, comp.graphics, sci.electronics	comp.sys.ibm.pc.hardware, sci.med, comp.sys.mac.hardware, comp.windows.x	0.75
问题 11	comp.graphics, sci.med, comp.windows.x, sci.space	comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, sci.crypt	0.75
问题 12	sci.crypt, talk.politics.guns, sci.electronics	sci.med, talk.politics.mideast, talk.politics.misc, talk.religion.misc	0.30