Journal of Computer Applications
计算机应用,2012,32(12):3308-3310
ISSN 1001-9081
CODEN JYIIDU
2012-12-01
http://www.joca.cn

文章编号:1001-9081(2012)12-3308-03
doi:10.3724/SP.J.1087.2012.03308

基于局部学习的半监督多标记分类算法

吕　佳
(重庆师范大学 计算机与信息科学学院,重庆 400047)
(*通信作者电子邮箱 lvjia@cqnu.edu.cn)

摘　要:针对在求解半监督多标记分类问题时通常将其分解成若干个单标记半监督二类分类问题从而导致忽视类别之间内在联系的问题,提出基于局部学习的半监督多标记分类方法。该方法避开了多个单标记半监督二类分类问题的求解,采用"整体法"的研究思路,利用基于图的方法,引入基于样本的局部学习正则项和基于类别的拉普拉斯正则项,构建了问题的正则化框架。实验结果表明,所提算法具有较高的查全率和查准率。

关键词:半监督学习;多标记分类问题;局部学习;标记;正则项
中图分类号:TP391.4　　**文献标志码**:A

Semi-supervised multi-label classification algorithm based on local learning

LÜ Jia*
(*College of Computer and Information Sciences, Chongqing Normal University, Chongqing* 400047, *China*)

Abstract: Semi-supervised multi-label classification problem is usually decomposed into a set of single-label semi-supervised binary classification problems. However, it results in the ignorance of the inner relationship between labels. A semi-supervised multi-label classification algorithm was presented, which avoided multiple single-label semi-supervised binary classification problems but adopted the overall approach in this paper. On the basis of undirected graph, local learning regularizer for data points and Laplace regularizer for labels were introduced and regularization framework of the problem was constructed. The experimental result shows the proposed algorithm has higher precision and recall.

Key words: semi-supervised learning; multi-label classification problem; local learning; label; regularizer

0　引言

多标记学习起源于文本分类研究中遇到的歧义性问题,主要解决一个样本可以同时属于多个类别的问题。现实世界中,多标记学习问题普遍存在[1],例如,在生物信息学中,一个基因序列具有若干个功能,如"新陈代谢"、"蛋白质合成"等;在文本分类中,每篇文档可能同时属于多个主题,如"苹果"、"乔布斯"等;在场景分类中,每个场景图片可能对应于多个类别,如"大海"、"沙滩"等。通常多标记分类问题的一种直观的处理办法是把多标记分类问题转化为一组独立的二类分类问题,其中每一个二类分类问题对应一个标记,每一个样本的标记最终通过组合所有的二类分类问题的结果获得[2]。这种处理方法的好处在于可以利用最新的二类分类算法,缺点是它是孤立地处理分解得到的每一个二类分类问题,未考虑到每个样本所属类标记集中类与类之间的相关性。而在实际问题中,如能充分利用类与类之间的相关性,则可有效地提高学习系统的泛化能力。研究者们已开始考虑样本所属类标记集中类与类之间的相关性的问题[3-4]。进一步地,在多标记分类问题中考虑无标记样本,这就是半监督多标记分类问题。将半监督学习引入多标记学习中,可以降低多标记学习在应用中的成本,使得仅需标注少量的样本,得到更好的学习效果。其中,比较具有代表性工作的有:Liu 等[5]提出了一种半监督多标记学习框架,根据输入空间和输出空间的相似性,把半监督多标记学习问题转换为约束非负矩阵因式分解的问题。该方法有效地利用了无标记样本提供的信息,并且考虑了类与类之间的相关性,在训练样本数相对较少时分类效果也很好。陈钢等[6]同时考虑无标记样本和类与类之间相关性两方面的内容,在训练样本和类标记上分别创建了无向图,构建了基于图的正则化框架。再通过求解 Sylvester 方程来获得无标记样本的标记。孔祥南等[7]利用直推式多标记分类(Transductive multi-label classification, TRAM)方法为每一个训练样本分配一组多标记,首先构造直推式多标记学习的优化问题来估计类标记构成,接着推导出该优化问题的闭型解,最后运用一种有效的算法给未标记样本分配标记集。本文以"整体法"来研究半监督多标记分类问题,利用局部学习来习得样本类标记,利用流形学习来考虑类别与类别之间的相关性,提出了基于局部学习的半监督多类分类算法,实验证明了算法的可行性和有效性。

1　半监督二分类问题

半监督多标记分类问题的数学描述如下:

给定训练集

$$T = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_l,\boldsymbol{y}_l),\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\} \tag{1}$$

其中:$\boldsymbol{x}_i \in \mathbf{R}^d (i = 1,2,\cdots,n)$ 称为输入,$\boldsymbol{y}_i \subseteq \bar{\boldsymbol{Y}} (i = 1,2,\cdots,l)$ 称为输出,$\bar{\boldsymbol{Y}} = \{1,2,\cdots,c\}$ 是 c 个标记构成的集合,即输出空间。试据此寻找与输入 $\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n$ 对应的在 $\bar{\boldsymbol{Y}}$ 中取值的输出 $\boldsymbol{y}_{l+1},\boldsymbol{y}_{l+2},\cdots,\boldsymbol{y}_n$ 的值。

在半监督多标记分类问题求解中,$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$ 转换为一个长度为 c 的由 0 和 1 构成的二进制序列

$$\hat{y}_{ik} = \begin{cases} 1, & k \in \boldsymbol{y}_i, k = 1,\cdots,c \\ 0, & \text{其他} \end{cases} \tag{2}$$

收稿日期:2012-07-25;**修回日期**:2012-08-29。
基金项目:国家自然科学基金资助项目(11071252);重庆市教委科技项目(KJ120628);重庆师范大学博士启动基金资助项目(12XLB030)。
作者简介:吕佳(1978-),女,四川达州人,副教授,博士,主要研究方向:机器学习、半监督学习。

$\boldsymbol{y}_{l+1},\boldsymbol{y}_{l+2},\cdots,\boldsymbol{y}_n$ 初始值转换为长度为 c 的全零向量。

$$\hat{\boldsymbol{y}}_i = (0,\cdots,0) \in \mathbf{R}^{1\times c} \tag{3}$$

这样,$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$ 就对应一个由 0 和 1 构成的 $n\times c$ 矩阵。计算得到的实值矩阵解 $\boldsymbol{F} = (\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n)^{\mathrm{T}} \in \mathbf{R}^{n\times c}$,$\boldsymbol{f}_i \in \mathbf{R}^c$,其中:$\boldsymbol{f}_i = (f_{i1},\cdots,f_{ic}) \in \mathbf{R}^{1\times c}$,$i = 1,2,\cdots,n$。输入 $\boldsymbol{x}_i$ 的输出 $\boldsymbol{y}_i$ $(i = 1,2,\cdots,n)$ 由式(4)确定:

$$\boldsymbol{y}_i = \{j \mid f_{ij} > s(\boldsymbol{x}_i), j \in \bar{\boldsymbol{Y}}\};\ i = l+1,\cdots,n \tag{4}$$

其中 $s(\cdot)$ 为对应的阈值函数。

2　基本思想

基于图的半监督分类算法的正则化学习框架[8-11]如下:

$$\min_{f\in H}\sum_{i=1}^{n} L(f(\boldsymbol{x}_i),\boldsymbol{y}_i) + \nu\psi(\boldsymbol{f}) \tag{5}$$

其中:H 是 Hilbert 空间,$\boldsymbol{f}$ 是实值函数,ν 是权衡系数。目标函数中第一项是经验损失,第二项是正则项。按照此框架,针对半监督多标记分类问题,除了损失项,需重点考虑如何构建目标函数中的正则项。

2.1　局部学习正则项

由 Vapnik[12-13]提出的局部学习算法,即一个样本的类别由其邻域内样本的类别学习得到,最初是针对有监督学习的,Wu 等[9]提出了局部学习正则项解决了局部学习算法在半监督二类分类问题中的应用。Xiang 等[11]采用二进制序列标记表示方法来表示半监督多类分类问题中的类标记。

$$L_{ij} = \begin{cases} 1, & \boldsymbol{y}_i = j \in \{1,2,\cdots,c\} \\ 0, & \text{其他} \end{cases} \tag{6}$$

这与半监督多标记分类问题中标记的表示方法类似,只不过前者的二进制序列中仅有其中一位为 1,其余位均为 0。

通过对局部学习模型分析发现,

$$\boldsymbol{G} = \boldsymbol{AF} \tag{7}$$

$\boldsymbol{A}$ 是独立于 $\boldsymbol{F}$ 的,只与输入 $\boldsymbol{x}_i$ $(i = 1,2,\cdots,n)$ 的邻域矩阵相关。这样就很自然地在半监督多标记分类问题中考虑局部学习的问题。而实际上,这里可以直接采用式(7)作为局部学习模型,详见文献[14]。

2.2　类与类之间相关性处理方法

局部学习正则项的作用是为了平滑训练样本,而训练样本所对应的标记集中各标记之间具有一定的相关性,蕴涵了类别的光滑性问题[6-7]。故对于标记的处理,仍然采用基于图的处理方法[6]。每一个输入 $\boldsymbol{x}_i$ 的输出 $\boldsymbol{y}_i$ $(i = 1,2,\cdots,n)$ 被转化为一个向量 $\hat{\boldsymbol{y}}_i \in \mathbf{R}^{1\times c}$ $(i = 1,2,\cdots,n)$,全体输入 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 对应的输出 $\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$ 被表示成一个 $n\times c$ 的矩阵,从每一个类别的角度来看待该矩阵,它就是由 c 个标记向量构成的。因此:

$$(\hat{\boldsymbol{y}}_1,\cdots,\hat{\boldsymbol{y}}_n)^{\mathrm{T}} = (\boldsymbol{t}_1,\cdots,\boldsymbol{t}_c) \tag{8}$$

其中:$\boldsymbol{t}_k \in \mathbf{R}^n$ $(k = 1,2,\cdots,c)$。

以 $\bar{\boldsymbol{Y}} = \{1,2,\cdots,c\}$ 中的 c 个标记构建一个加权无向图,图中的每一个顶点对应一个标记,连接顶点与顶点之间的边的权值由式(9)得到:

$$W_{ij} = \exp\left(-\frac{1}{\delta}(1-\cos(\boldsymbol{t}_i,\boldsymbol{t}_j))\right) \tag{9}$$

其中:$\delta > 0$ 是一个超参数,$\cos(\boldsymbol{t}_i,\boldsymbol{t}_j)$ 用来度量顶点 $\boldsymbol{t}_i$ 和顶点 $\boldsymbol{t}_j$ 之间的相似度:

$$\cos(\boldsymbol{t}_i,\boldsymbol{t}_j) = \frac{\langle \boldsymbol{t}_i,\boldsymbol{t}_j\rangle}{\|\boldsymbol{t}_i\|_2\|\boldsymbol{t}_j\|_2} \tag{10}$$

其中:$\langle \boldsymbol{t}_i,\boldsymbol{t}_j\rangle$ 是 $\boldsymbol{t}_i$ 和 $\boldsymbol{t}_j$ 的内积,$\|\cdot\|_2$ 是欧式向量范数,相似度 $\cos(\boldsymbol{t}_i,\boldsymbol{t}_j) \in [0,1]$,相似度越接近于 1,$\boldsymbol{t}_i$ 和 $\boldsymbol{t}_j$ 越相似,反之亦然。

半监督多标记分类问题中,求解出来的实值矩阵解 $\boldsymbol{F}$ 表示为:

$$\boldsymbol{F} = (\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n)^{\mathrm{T}} \in \mathbf{R}^{n\times c} = (\boldsymbol{p}_1,\cdots,\boldsymbol{p}_c) \tag{11}$$

其中:$\boldsymbol{f}_i$ $(i = 1,2,\cdots,n)$ 是从样本的角度考察得到的 c 维实值解,令 $\boldsymbol{P} = (\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_c)^{\mathrm{T}}$,$\boldsymbol{p}_i$ $(i = 1,2,\cdots,c)$ 是从标记的角度考察得到的 n 维实值解,且 $\boldsymbol{F} = \boldsymbol{P}^{\mathrm{T}}$。则根据流形假设[8],希望 $\psi(\boldsymbol{P}) = \frac{1}{2}\sum_{i,j=1}^{c} W_{ij}\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2$ 尽可能小,以此满足无向图上的顶点之间的光滑性。

通过变换,可得到 $\psi(\boldsymbol{P})$ 的矩阵表达形式

$$\psi(\boldsymbol{P}) = \frac{1}{2}\sum_{i,j=1}^{c} W_{ij}\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2 =$$

$$\frac{1}{2}\sum_{i,j=1}^{c} W_{ij}(\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i - 2\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_j + \boldsymbol{p}_j^{\mathrm{T}}\boldsymbol{p}_j) =$$

$$\frac{1}{2}\Big(\sum_{i=1}^{c} D_{ii}\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i - 2\sum_{i,j=1}^{c} W_{ij}\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_j + \sum_{i=1}^{c} D_{jj}\boldsymbol{p}_j^{\mathrm{T}}\boldsymbol{p}_j\Big) =$$

$$\mathrm{trace}(\boldsymbol{P}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{P}) = \mathrm{trace}(\boldsymbol{F}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{F}^{\mathrm{T}}) = \mathrm{trace}(\boldsymbol{FLF}^{\mathrm{T}})$$

令

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W} \tag{12}$$

其中:$\boldsymbol{W} = (w_{ij}) \in \mathbf{R}^{n\times n}$ 是一个对称半正定矩阵;$\boldsymbol{D}$ 是一个对角矩阵,其对角元素 $D_{ii} = \sum_{j=1}^{n} w_{ij}$ $(i = 1,2,\cdots,n)$;$\boldsymbol{L}$ 即为拉普拉斯正则项因子。

故解决类与类之间相关性的问题就转化为以下的二次优化问题

$$\min_{\boldsymbol{F}\in\mathbf{R}^{n\times c}} \boldsymbol{FLF}^{\mathrm{T}} \tag{13}$$

3　基于局部学习的半监督多标记分类算法

按照正则化学习框架(式(5)),可得到基于局部学习的半监督多标记分类问题:

$$\min_{\boldsymbol{F}\in\mathbf{R}^{n\times c}} (\boldsymbol{F}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{F}-\boldsymbol{Y}) + \boldsymbol{F}^{\mathrm{T}}\boldsymbol{OF} + \gamma_I\boldsymbol{FLF}^{\mathrm{T}} \tag{14}$$

其中:$\gamma_I > 0$ 是一个权衡系数。第一项是经验风险,最小化其值以使最优解 $\boldsymbol{F}$ 尽可能与实际类标记一致;第二项是基于训练样本的局部学习正则项,最小化该项以使 $\boldsymbol{F}$ 具有希望的理想性质;第三项是基于标记的拉普拉斯正则项,最小化该项以使标记尽可能光滑。

上述问题是一个无约束最优化问题,通过向量求导的方法求解,得到

$$(\boldsymbol{O}+\boldsymbol{D})\boldsymbol{F} + \boldsymbol{FL} = \boldsymbol{DY} \tag{15}$$

其中:$\boldsymbol{O} \in \mathbf{R}^{n\times n}$,$\boldsymbol{D} \in \mathbf{R}^{n\times n}$,$\boldsymbol{F} \in \mathbf{R}^{n\times c}$,$\boldsymbol{L} \in \mathbf{R}^{c\times c}$,$\boldsymbol{Y} \in \mathbf{R}^{n\times c}$。而这正是在控制论和通信理论中起着重要作用的 Sylvester 方程。

许多文献都在讨论 Sylvester 方程的解法,其中文献[15]针对广义的 Sylvester 方程,提出了利用 Galerkin 和最小残差算法来迭代求解。本文直接利用该算法来求解最优化问题(式(14))的最优解 $\boldsymbol{F}$。

下面给出基于局部学习的半监督多标记分类(semi-supervised Multi-Label Classification algorithm based on Local Learning, LL_MLC)算法的详细流程:

1)给定如式(1)所示的训练集。

2)选择适当的参数 $D_l > 0$,$D_u \geqslant 0$,$\lambda > 0$,$\gamma_I > 0$,$\delta > 0$,

阈值函数 $s(\cdot)$,最近邻样本数 $K > 0$。

3)根据式(2)和(3)创建 $n \times c$ 矩阵 $\boldsymbol{Y}$。

4)计算 $\boldsymbol{x}_i$ 的最近 K 个输入。

5)按照文献[14]中的方法创建矩阵 $\boldsymbol{A}$。

6)根据式(12)计算 L。

7)按照文献[15]中求解 Sylvester 方程的算法计算 $\boldsymbol{F}$。

8)根据式(4)确定 $\boldsymbol{y}_{l+1},\cdots,\boldsymbol{y}_n$。

4 实验结果

4.1 实验数据集

实验数据集采用目前应用最为广泛的文本分类基准数据集 Reuters 数据集中的 Reuters-21578 Distribution 1.0 版本。该数据集中包含 21 578 个文本,其中仅有不到一半的文本具有人工赋予的标记。在有标记样本中随机选择 3 000 个文本作为初始数据集,每个文本对应于 98 个类中的若干个类组合。实验中随机选取初始数据集中的 500 个样本作为有标记样本,忽略剩下的 2 500 个有标记样本的标记,将它们视为无标记样本,从而得到本文所需的实验数据集。

4.2 实验方法与参数设置

半监督多标记分类中采用文献[6]中的 $MicroF_1$ 作为评估标准。第 k 类的 $MicroF_1$ 的度量方式为

$$MicroF_1(k) = \frac{2p_k r_k}{p_k + r_k} \tag{16}$$

其中:p_k 是第 k 类的查准率(precision),r_k 是第 k 类的查全率(recall)。$MicroF_1$ 则综合了查准率和查全率,将二者赋予同样的重要性来考虑。$MicroF_1$ 的取值在 0 和 1 之间,其值越接近于 1,系统性能越好。p_k 和 r_k 的计算方法如下:

$$p_k = \frac{|\{\boldsymbol{x}_i \mid k \in \boldsymbol{y}_i \wedge k \in \bar{\boldsymbol{y}}_i\}|}{|\{\boldsymbol{x}_i \mid k \in \bar{\boldsymbol{y}}_i\}|} \tag{17}$$

$$r_k = \frac{|\{\boldsymbol{x}_i \mid k \in \boldsymbol{y}_i \wedge k \in \bar{\boldsymbol{y}}_i\}|}{|\{\boldsymbol{x}_i \mid k \in \boldsymbol{y}_i\}|} \tag{18}$$

其中:$\boldsymbol{y}_i$ 是样本的真实标记,$\bar{\boldsymbol{y}}_i$ 是样本经学习后得到的标记。

为了验证算法的有效性,将 LL_MLC 算法同文献[5]提出的约束非负矩阵分解的半监督多标记学习算法(Semi-supervised multi-label learning by constrained non-negative matrix factorization, CNMF)以及文献[6]提出的求解 Sylvester 方程的半监督多标记学习算法(Semi-supervised algorithm for multi-label learning by solving a Sylvester equation, SMSE)做了对比实验。CNMF 算法根据输入空间和输出空间的相似性,把半监督多标记学习问题转换为约束非负矩阵因式分解的问题。SMSE 算法首先在训练样本和类别标记上分别创建了连通加权无向图,构建了基于图的正则化框架,再通过求解 Sylvester 方程来获得未标记样本的标记。需要说明的是,该算法虽然最终也要求解一个 Sylvester 方程,但与本文算法不同,它在训练样本上得到的正则项是基于流形假设,而本文的正则项是基于局部假设。此外,Sylvester 方程的求解方法采用的是求解广义 Sylvester 方程的一种优良的算法,该算法已证明具有较好的性能。

所有算法的最优参数都是通过最小化留一法(Leave-One-Out, LOO)进行网格搜索得到,参数设置情况如下:$\lambda \in \{0.1,1,10,100\}$,$K \in \{5,10,20,50,100\}$,$D_l \in \{0.1,1,10,100\}$,$D_u = 10^{-6}$,$\gamma_I \in \{2^{-5},\cdots,2^5\}$,$\delta \in \{2^{-5},\cdots,2^5\}$。

4.3 实验结果

实验重复 20 次,得到三种算法在实验数据集上的平均 $MicroF_1$ 值,见图 1。从图中可以看出,LL_MLC 算法要优于 CNMF 算法和 SMSE 算法。

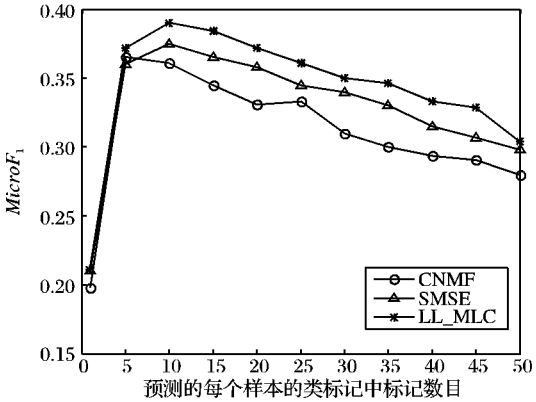

图 1 三种算法的平均 $MicroF_1$ 值曲线

5 结语

本文分析了半监督多标记分类问题中类标号表示的特点,根据它与半监督多类分类问题中二进制序列标记表示方法的一致性,将局部学习正则项引入到半监督多标记分类问题中。其次考虑了类别之间相关性的影响,针对类别构建了基于图的拉普拉斯正则项,从而提出了基于局部学习的半监督多标记分类算法。最后通过实验验证了算法是可行的。

在多标记学习问题中,标记之间的相关性作为整体研究是通过图的拉普拉斯正则化的方式来反映的,而若分解成单标记问题,可以利用关联规则,这实际上涉及到的就是关系学习理论[16],若仍然采用"整体法"的研究思路,将多标记学习和关系学习相互结合,不仅可以为多标记学习提供新的研究思路,还可以拓展关系学习理论的研究空间。

参考文献:

[1] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.

[2] CHANG E, GOH K, SYCHAY G, *et al*. Content-based soft annotation for multimodal image retrieval using Bayes point machines[J]. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description, 2003, 13(1): 26-38.

[3] BOUTELL M R, LUO J SHEN X, *et al*. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[4] ZHANG MINLING, ZHANG KUN. Multi-label learning by exploiting label dependency[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 999-1007.

[5] LIU YI, JIN RONG, YANG LIU. Semi-supervised multi-label learning by constrained non-negative matrix factorization[C]// Proceedings of the 21st National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2006: 421-426.

[6] CHEN GANG, SONG YANQIU, WANG FEI, *et al*. Semi-supervised multi-label learning by solving a Sylvester equation[EB/OL]. [2012-05-20]. https://www.siam.org/proceedings/datamining/2008/dm08_37_chen.pdf.

[7] KONG X N, NG M K, ZHOU Z H. Transductive multi-label learning via label set propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 788-799.

(下转第 3338 页)

表2 对语料B的分类实验结果对比 %

类别	对比实验			本文实验方法1			本文实验方法2		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
第一类: culture	55	66.667	60.274	52.5	87.500	65.625	90.0	78.261	83.721
第二类: military	75	88.235	81.081	80.0	84.211	82.052	95.0	79.167	86.364
第三类: reading	65	65.000	65.000	70.0	66.667	68.291	80.0	94.118	86.487
第四类: society&law	85	64.151	73.118	92.5	66.071	77.083	77.5	96.875	86.111
平均	70	71.013	70.503	73.75	76.112	74.912	85.625	87.105	86.359

从语料A的实验结果可以看到,除了对第三类的分类精度本文实验方法2小于对比实验外,其他类别的精度都是本文方法的分类精度相对较高,同时也注意到本文方法的实验在第三类的召回率、*F*1值相对对比实验是较高的。同时观察语料可以发现,由于语料A取自百度知道这样一个网友自发组织内容的平台,其中的问题语料难免含有大量语义信息不足的符号,比如"病了怎么办?????????",类似这样的预测语料分词并去除停用词之后,由于关键词相当少,因此可能使用对比实验的方法效果更加不理想,而本文方法对其进行概念描述后,再通过在训练语料中寻找概念层上相同的语料进行扩展后,更加有助于提高分类的效果。对于语料B的实验结果,除了第一类的分类精度本文实验方法1小于对比实验外,其他类别的精度都是本文方法的实验结果精度相对较高。

总的来看,使用本文方法将短文本分类问题转换成长文本分类问题后,不管是使用朴素贝叶斯分类器还是支持向量机分类器,都相对原来直接使用朴素贝叶斯分类器在短文本上进行分类的效果有所提升。因为短文本虽文本长度不长,但是同样要表达一个意思,短文本中每个词包含的语义信息相对更多,针对这个特点充分挖掘短文本每个词背后的语义信息,使得算法的分类效果更佳。

3 结语

本文提出了一种高效并且易于扩展的短文本算法,该算法利用短文本信息短因此其中每个词包含的信息量更多的特点,将短文本在概念层上进行一个扩展,再使用已经相对成熟的长文本分类算法对短文本进行分类,实验证明比传统的分类方式正确率更高。在以后的工作中,将研究噪声对实验的影响,进一步提高算法的精度和适用性。

参考文献:

[1] SIMM W, FERRARIO M A, PIAO S, *et al*. Classification of short text comments by sentiment and actionability[C]// IEEE International Conference Social Computing/IEEE International Conference on Privacy, Security. Washington, DC: IEEE Computer Society, 2010: 552-557.

[2] 蔡月红, 朱倩, 孙萍, 等. 基于属性选择的半监督短文本分类算法[J]. 计算机应用, 2010, 30(4): 1015-1018.

[3] 王永恒, 贾焰, 杨树强. 大规模文本数据库中的短文本分类方法[J]. 计算机工程与应用, 2006,42(22): 5-7.

[4] HEALY M, DELANY S J, ZAMOLOTSKIKH A. An assessment of case base reasoning for short text message classification[C]// Proceedings of the 16th Irish Conference on Artificial Intelligence and Cognitive Science. Ireland: Dublin Institute of Technology, 2005: 1-10.

[5] 樊兴华, 王鹏. 基于两步策略的中文短文本分类研究[J]. 大连海事大学学报, 2008, 34(3): 121-124.

[6] 郭泗辉, 樊兴华. 一种改进的贝叶斯网络短文本分类算法[J]. 广西师范大学学报: 自然科学版, 2010, 28(3): 140-143.

[7] 林小俊, 张猛. 基于概念网络的短文本分类方法[J]. 计算机工程, 2010, 36(21): 4-6.

[8] 刘伍颖, 王挺. 基于词模型索引的短文本在线过滤方法[J]. 华中科技大学学报: 自然科学版, 2010, 38(4): 42-45.

[9] ZELIKOVITZ S, HIRSH H. Improving short-text classification using unlabeled background knowledge to assess document similarity[C]// Proceedings of the Seventeenth International Conference on Machine Learning. New York: Morgan Kaufmann, 2000: 1191-1198.

[10] SRIRAM B, FUHRY D, DEMIR E. Short text classification in twitter to improve information filtering[C]// Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2010: 841-842.

[11] 王细薇, 樊兴华. 一种基于特征扩展的中文短文本分类方法[J]. 计算机应用, 2009, 29(3): 843-845.

[12] 汉语国际教育技术研发中心[EB/OL]. [2011-10-27]. http://nlp.blcu.edu.cn/downloads/download-resources/36.html.

[13] 闫瑞, 曹先彬, 李凯. 面向短文本的动态组合分类算法[J]. 电子学报, 2009, 37(5): 1019-1024.

(上接第3310页)

[8] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006, 7(11): 2399-2434.

[9] WU M R, SCHOLKOPF B. Transductive classification via local learning regularization[C]// Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. Cambridge: MIT Press, 2007: 624-631.

[10] WANG FEI. A general learning framework using local and global regularization[J]. Pattern Recognition, 2010, 43(9): 3120-3129.

[11] XIANG SHIMING, NIE FEIPING, ZHANG CHANGSHUI. Semi-supervised classification via local spline regression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 2039-2053.

[12] VAPNIK V. The nature of statistical learning theory[M]. Berlin: Springer-Verlag, 1995.

[13] BOTTOU L, VAPNIK V. Local learning algorithms[J]. Neural Computation, 1992, 4(6): 888-900.

[14] 吕佳. 结合全局和局部正则化的半监督二分类算法[J]. 计算机应用, 2012, 32(3): 643-645, 648.

[15] BAO LIANG, LIN YIQIN, WEI YIMIN. Krylov-subspace methods for the generalized Sylvester equation[J]. Applied Mathematics and Computation, 2006, 175(1): 557-573.

[16] DZEROSKI S, LAVRAC N. Relational data mining[M]. Berlin: Springer, 2001.