文章编号:1000-5862(2014)01-0102-06

基于 Markov 随机游走的渐进式半监督分类模型

陈秀平[1],王明文[1*],万剑怡[1],左家莉[2]
(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 江西师范大学初等教育学院,江西 南昌 330027)

摘要:提出了一种基于 Markov 随机游走的渐进式半监督分类模型:在随机游走过程中,计算待标注数据到各类的迁移概率时,只考虑相应类别样本的影响,而忽略其他类别样本对随机过程的影响;并在学习过程中借鉴渐进学习思想,通过不断地"纠正"半监督学习过程中的"错误",从而提高模型的预测精度. 在 20newsgroups 数据集上的实验结果表明:所提出的方法能够提高半监督分类的精度.

关键词:半监督分类;渐进学习;Markov 随机游走;迭代
中图分类号:TP 311 **文献标志码**:A

0 引言

随着互联网技术的迅猛发展,信息量迅速增加,未标记数据随手可得,而大量样本的人工标注往往需要耗费大量的人力和物力,且要求标注者具有较强领域的专业知识,使得付出的代价过于"昂贵"[1]. 这给信息的检索和获取带来了巨大的困难,如何在浩瀚的信息海洋中获取需要的信息是目前亟需解决的问题. 若能只对少量的数据进行标注,就可以实现对信息进行有效地组织和管理,将是非常有意义的.

传统的监督学习往往需要足够多的已标记的学习样本进行训练,才能获得具有较好的泛化性能的监督学习方法. 而在实际应用中,要获得有标记的样本相当困难. 在通常情况下无监督学习,试图利用未标记样本的隐含信息来获得相应的学习器,将很难保证较高的学习精度[2]. 半监督学习作为近年来的研究热点,能够同时利用标记数据和未标记数据进行学习,弥补了有监督学习与无监督学习的不足,适合于已标记样本较少,同时具有大量未标记样本的分类问题[3]. 但由于半监督学习只利用少量的样本信息,分类效果并不理想.

本文针对如何综合利用未标记和标记样本信息以提高半监督分类性能这一问题,提出了一种基于 Markov 随机游走的渐进式半监督分类模型. 该算法首先利用简单的 k-means 算法对数据进行初始聚类,保留聚类后每个簇的簇中心,然后选择离簇中心最近的几个点进行标注,作为 Markov 随机游走的起始点;在游走的过程中,利用渐进学习的思想,通过更新 Markov 随机游走图中节点间的迁移概率权重,不断"纠正"半监督学习过程中的"错误",从而提高半监督分类算法的精度. 在 20newsgroups[4] 数据集上的实验结果表明,该算法可以有效地提高基于 Markov 随机游走的半监督分类预测精度.

1 相关工作

半监督学习作为一种能综合利用少量标注样本数据和大量未标注样本数据来提高学习性能的学习方法,已成为机器学习领域中的一个研究热点. 目前,半监督学习算法主要有:(ⅰ)基于图的模型(graph-based model)的半监督学习;(ⅱ)基于生成式混合模型(generative mixture model)的半监督学习;(ⅲ)直推式支持向量(transductive SVM)的半监督学习;(ⅳ)基于协同训练(co-training)模型的半监督学习[1].

有很多学者对利用少量标记数据进行学习的半监督文本分类进行了广泛研究:Liu Bing 等[5] 提出了 S-EM 算法用于监督的文本分类,郑海清等[6] 提出了一种基于紧密度衡量的半监督文本分类算法,Blum 等[7] 利用协同训练算法研究了 Web 网页分类

收稿日期:2013-11-17
基金项目:国家自然科学基金(60963014)资助项目.
通信作者:王明文(1965-),男,江西南康人,教授,博士生导师,主要从事信息检索和并行计算的研究.

问题. M. Szummer 等[8-12]提出了在 Markov 随机游走图上利用少量标记数据进行文本分类算法,文中在 Markov 随机游走图使用 EM 算法来估计分类边界,从而达到半监督分类的效果. 该方法未考虑在随机游走的过程中,当经过不同类别文档时加入的错误分类信息会影响分类精度的情况.

2　基于 Markov 随机游走的渐进式半监督分类模型

2.1　Markov 随机游走图的生成

将训练集合 X 映射成多维度空间中的随机游走图. 在 Markov 随机游走模型中,通过点与点之间的连通性可以更好地刻画训练数据之间的相关性,其基本思路是将集合 X 中的每个训练数据 $x_i \in X$ 映射为图中的一个点 y_i,训练集合 X 对应的随机游走图可表示为

$$G(V,E)\,, \tag{1}$$

$$V = \{v_i \mid x_i \in X, 1 \leqslant i \leqslant n\}\,, \tag{2}$$

$$E = \{(u,v) \mid u,v \in V\}\,, \tag{3}$$

其中 V 为顶点的集合,E 为节点间的相互关系,v_i 为每个训练数据 x_i 在 Markov 随机游走图上对应的顶点. 如(2)式表示每个训练数据 x_i 对应图 G 中的顶点 v_i,这些顶点构成了图的顶点集合 V.

计算 Markov 随机游走图 G 上的权重矩阵 W,用欧式距离作为距离函数计算训练数据间的空间距离,记作 $d(x_i,x_j)$,图 G 的边权值为 $W_{ij} = \exp(-d(x_i,x_j)/\sigma^2)$,从而构建了一个以 W_{ij} 为边权的无向图. 对图 G 取其 k 邻近,得到所需要的图 G'. 在图 G' 中从点 i 到点 j 的一步迁移概率 p_{ij} 可以表示为

$$p_{ij} = W_{ij} \Big/ \sum_k W_{ik}. \tag{4}$$

根据(4)式可以得到 Markov 随机游走转移概率矩阵 P. 很多半监督的分类模型都是基于该表示方法[5]. $P_{ij} = 0$ 表示节点 j 没有邻居,W_{ij} 是对称的,但 p_{ij} 通常是不对称的,这是对每一个节点归一化的结果. $P_{0|t}(i|j)$ 表示从节点 i 到节点 j 的 t 步迁移概率(这里 t 是一个参数,而不是一个随机变量),若将一步迁移概率表示为矩阵 A,则可以用矩阵乘法来计算 t 步迁移概率为

$$P_{0|t}(i/j) = [A^t]_{ij}\,, \tag{5}$$

因为矩阵 A 中每一行是服从随机分布的,所以矩阵 A 的行和为 1.

2.2　基于 Markov 随机游走半监督分类模型

基于 Markov 随机游走的半监督分类模型(SMRW)的思想是:(ⅰ)通过 k-means 聚类算法从未标记的文档集合中选择可信的、典型的样例进行标注;(ⅱ)从已标注的样例出发,在 Markov 随机游走图上进行游走,计算某一待标注样本 j 在 t 步内到达正例和负例的迁移概率分别为 p_j^+、p_j^-,其中计算 p_j^+ 时不经过已标注为负类的样例,对于 p_j^- 则反之. 比较 p_j^+ 与 p_j^-,若 $p_j^+ > p_j^-$,则将 j 标注为正类,否则标注为负类.

更一般地,若将数据集表示为 $X = X_L + X_U = \{x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_n\}$,其中 X_L 表示标记样本,X_U 表示未标记样本,$Y = (y_1,y_2,\cdots,y_c)$ 是对应的类标集合,c 表示类别数. X_L 中每类标记文档数为 m,则可将在 Markov 网络中的随机游走描述为

(ⅰ)当 $t = 1$ 时,

$$p_{0|1}(Y|j) = p_{t=1}(Y|j) = \sum_{i=1}^{m} p(\mathrm{Node}(i)\mid \mathrm{New}) \cdot \mathrm{e}^{-\alpha\cdot 1}; \tag{6}$$

(ⅱ)当 $t = 2$ 时,

$$p_{0|2}(Y\mid j) = p_{t=1}(Y|j) + p_{t=2}(Y|j) = \sum_{i=1}^{m} p(Node(i)\mid New) \cdot \mathrm{e}^{-\alpha\cdot 1} + \sum_{i=1}^{m} p(Node(i)\mid New) \cdot \mathrm{e}^{-\alpha\cdot 2}; \tag{7}$$

(ⅲ)更一般地,

$$p_{0|t}(Y|j) = \sum_{t=1}^{T}\sum_{i=1}^{m} p_{t=i}(Node(i)\mid New) \cdot \mathrm{e}^{-\alpha\cdot t}, \tag{8}$$

其中 $\mathrm{e}^{-\alpha\cdot t}$ 为衰减因子,New 为待标注的点,Node(i)为 m 个已经标注点中的第 i 个节点,t 为当前游走步数,T 为总的游走步数.

根据(8)式,若要计算节点 j 属于某一类别 y_i 的概率可以通过计算从属于类别 y_i 每一个节点出发,在 t 步之内迁移到节点 j 的迁移概率来得到,那么节点 j 的最终类标通过 $C_j = \arg\max_{1\leqslant i\leqslant c} p(Y = y_i|j)$ 来确定,即将 t 步之内到节点 j 的迁移概率最大类的类标作为 j 的标签. 具体算法可描述为:

(A)输入:待分类的文档集合 $X = (x_1,x_2,\cdots,x_n)$;

(B)输出:$X \rightarrow Y$;

(C)算法步骤:

(ⅰ)对待分类文档进行 k-means 聚类,同时记录

簇中心信息,计算各文档到簇中心的距离 $d(x_i, cluster_c)$,并对各簇中的 $d(x_i, cluster_c)$ 进行降序排序,选择前 m 个文档进行标注,作为 Markov 随机游走的起始点;

(ii) 以已标注点作为随机游走的起始点,开始在 Markov 网络中进行 t 步游走,计算每个点到各类已分类文档的 t 步迁移概率 $p_{0|t}(Y|j) = \sum_{t=1}^{T}\sum_{i=1}^{m} p_{t=i}(Node(i)|New)\,e^{-\alpha\cdot t}$,并按 $C_j = \arg\max_{1\leqslant i\leqslant c} p(Y = y_i|j)$ 策略来标注该文档类标;

(iii) 重复以上步骤,直到该 Markov 网络中所有节点都成功标注.

该算法在进行游走的过程中,计算节点 j 到某一类别 y_i 的迁移概率时,迁移过程是不经过已标注为其它类别的节点的,并通过衰减因子 $e^{-\alpha\cdot t}$ 来约束不同迁移步数对迁移概率的影响. 实验过程中发现,SMRW 模型在迭代的过程中对样本的错分导致的误差,容易在之后的迭代中被不断放大,使得 SMRW 分类精度不高.

2.3 加入渐进学习思想的半监督 Markov 随机游走

基于 Markov 随机游走的渐进式半监督分类模型(PSMRW)的主要思想是:在已分类的各类文档中分别选择 n 篇文档,计算这 n 篇文档的错误率 ε,对错分的文档进行惩罚,即加大错分文档与其相同类标文档之间的迁移概率权重,而降低与其不同类标文档间的迁移概率权重,依次进行下去,使得分类中的"错误"得到"纠正",从而达到提高学习效率的目的. 具体算法可描述为

输入:$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其中,$x_i \in X$,$y_i \in Y = \{+1,-1\}$;初始迁移矩阵:$\boldsymbol{P}_1$;

(i) 初始化:$W_1(i) = 1/n$;

for $t = 1$ to T // T 为迭代次数

在 W_t 下进行半监督 Markov 随机游走的分类,得到弱学习分类结果:$h_t: X \to \{+1,-1\}$;计算 h_t 的错误率:$\varepsilon_i = \sum W_t(i)$,当 $h_t(x_i) \neq Y_i$ 时,令 $\alpha_t = \ln(1-\varepsilon_t)/\varepsilon_t)/2$;

(ii) 更新:对于样本 i,$h_t(x_i) \neq Y_i$,使其与相同类别样本间的迁移概率权重更新为 $W_{t+1}(i) = W_t(i)/Z_i \cdot e^{\alpha_t}$;与不同类别样本间的迁移概率权重更新为 $W_{t+1}(i) = W_t(i)/Z_t \cdot e^{-\alpha_t}$;$\boldsymbol{P}_{t+1} = \boldsymbol{P}_t \cdot W_{t+1}$;循环结束,输出 $H(X)$. 其中 z_t 为归一化因子 $z_t = \sum_{i=1}^{n} W_t(i)$,算法中 $h_1(t=1)$ 是一个弱学习器,是指其准确率仅比随机猜测的学习算法略高,即 $\varepsilon_t \leqslant 0.5$,本算法通过更新文本之间的迁移概率权重来指引分类器正确分类. 因为 $\varepsilon_t \leqslant 0.5$,采用 $\alpha_t = \ln((1-\varepsilon_t)/\varepsilon_t)/2$ 可得 $\alpha_t \geqslant 0$,从而使得对于误分的样本,其与相同类别样本之间迁移概率权重 $W_{t+1}(i) = e^{\alpha_t} W_t(i)/Z_t$ 加大,而其与不同类别样本之间的迁移概率权重 $W_{t+1}(i) = e^{-\alpha_t} W_t(i)/Z_t$ 减小,这样在之后的迭代中该样本被正确分类的概率增加.

3 实验设计及结果分析

3.1 实验准备

本次实验中,使用了文本分类中常用的评价数据集 20 newsgroups(20 个新闻组,英文数据集)进行实验.

在进行实验前,首先需要对文档数据进行预处理,主要进行了如下处理:(i) 去除文档中的格式标记、过滤非法字符、字母大小写转化、去除停用词等;(ii) 利用 DF(文档频率)进行特征选择,同时删除 DF 小于 3 的词项;(iii) 词干化,利用 Martin Porter 所提出的 Porter Stemmer 算法进行词干化处理;(iv) 采用 LTC 权重公式计算文档中词项的权重.

3.2 实验设计

3.2.1 *初始点的选择* 对于初始点的标注,不少研究者采用的是随机选择样本点的方法[2]. 每个样本点被选中标注为起始点的概率相等,这种标注样本选择法简单易行,被广泛使用,但收敛速度较慢.

本文中利用快速、高效的 k-means 算法对训练文档聚类,进行多次试验,计算其聚类纯度(纯度是一个评价聚类后每个类包含原始类某个类中的文档数目的指标),选择聚类纯度较高的聚类结果(试验中选取的聚类纯度为 0.9),保存其各簇中心信息. 对于每个簇,选择离簇中心最近的 m 个点标注类标,可选择可信、典型的样本进行标注.

3.2.2 *对比试验设计* 本文中主要进行了 2 组试验,一组是验证 SMRW 分类性能的实验,另一组是验证加入渐进学习思想的 PSMRW 的分类实验. 本文中分类使用的评价指标是错误率和 F_1 测度值,F_1 能综合考虑准确率和召回率.

3.3 参数调整

模型中有 3 个重要参数需要选择:剪枝邻接点

数 k,衰减函数参数 α 和最优游走步数 t.

3.3.1　剪枝邻接点数 k 的选择　在 Markov 随机游走模型中,每个节点的邻接点数 k 值的大小影响着模型的时间复杂度和精确度,k 值太小,容易造成随机游走图不连通,k 值太大易引入不大相关的点,从而降低模型精度和加大模型的复杂度. 本实验中,在 Windows 和 Mac 新闻组数据上取 $\sigma = 2$,初始标记点数 $n_l = 128$,$\alpha = 20$ 的情况下,分别比较了 k 从 1 到 50 的情况下,分类得到的 F_1 值结果如图 1 所示. 从

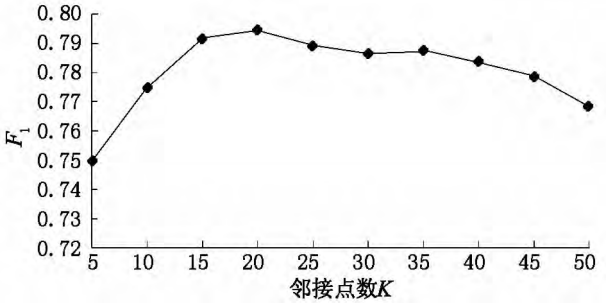

图 1　不同 k 值下的 F_1 值比较

图 1 中可以发现,当 $k = 20$ 时,分类的效果最佳,在以后的实验中,选取了 $k = 20$ 作为邻接点的个数.

3.3.2　衰减函数参数 α 值的选择　在 Markov 随机游走模型中,通过累加各步转移概率对文档间的间接相关性进行描述,通过不同的 α 来调整不同步转移概率的重要程度. 随着 α 值的增加,低步数概率的重要性不断增加. 为了确定 α 的最优值,在 Windows 和 Mac 新闻组数据上取 $k = 10$,$\sigma = 2$,$n_1 = 128$,分别比较了 α 从 1 到 100 的情况下,得到的 F_1 值结果如图 2 所示. 从图 2 中可以发现,当 $\alpha = 20$ 时,分类效果最佳,所以在模型中,选择 $\alpha = 20$ 作为衰减函数中的参数值.

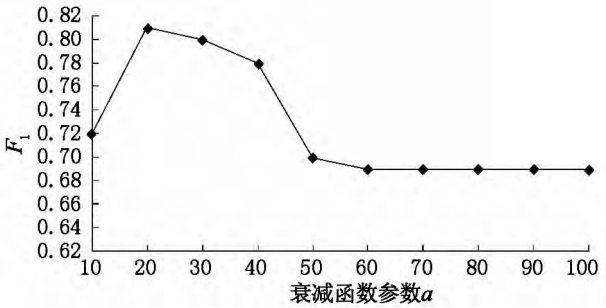

图 2　不同 α 值下的 F_1 值比较

3.3.3　最优游走步数 t 的选择　随机游走步数 t 决定了文档之间的间接关系的转移步数,在保证时间和空间复杂度一定的情况下,为了找到最能描述文档间相关性的最佳的随机游走步数,在 Windows 和 Mac 数据集上,选取 $\sigma = 2$,$k = 10$,$\alpha = 20$,$n_1 = 128$ 的情况下,分别比较了 t 从 1 到 30 变化过程中的分类 F_1 值的变化,试验中发现 $t = 10$ 之后,F_1 值基本保持不变,因此图 3 中只给出了 t 从 1 到 10 时 F_1 值的变化情况. 从图 4 中发现当 $t = 4$ 时,分类结果最佳. 结合图 4,综合考虑时间复杂度和分类精度,在今后的实验中,选取了 $t = 4$ 作为最高随机游走步数.

3.3.4　实验结果及分析　本实验采用的数据集是 20newsgroups 中的 Windows 和 Mac 数据,该数据包含 2 个类别,其中 Windows 类包含 985 篇,Mac 类包含 963 篇,共 1 948 篇.

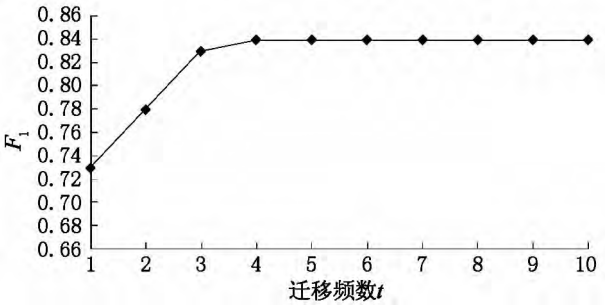

图 3　不同 t 值下的 $F1$ 值比较

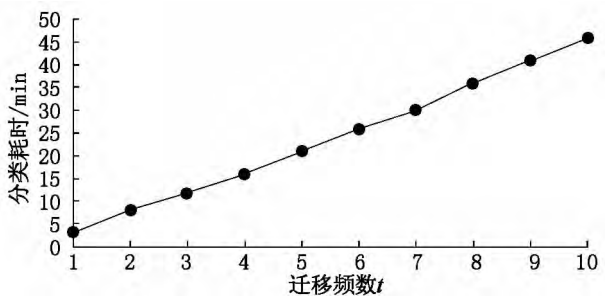

图 4　不同 t 值下的分类耗时比较

为了验证本文所提出的基于 Markov 随机游走的渐进式半监督分类模型的分类精度,进行了如下实验:(ⅰ) 不同初始标注文档数下,随机游走步数 t 对 F_1 的影响;(ⅱ) SMRW 与 *Mowkov* 平均边际分类方法(MAN) 和半监督支持向量机(semi-SVM) 方法的对比实验;(ⅲ) SMRW 和 PSMRW 2 种方法的对比实验; 实验中选定 $t = 4$,$\alpha = 20$,$k = 20$,$\sigma = 2$. 图 5 为 SMRW 方法分类的结果,横轴是随机行走的步数 t,纵轴是 F_1 值,从下往上,初始点个数为 2,8,64,128. 在已知数据非常稀少的情况下,分类的 F_1 值随着随机行走步数的增加首先降低,然后升高,然而当已知数据逐渐增多的时候,这种现象就不再发生了,分类的 F_1 值逐渐升到一个固定值. 这是因为当标记的初始点数较少时,可用的信息较少,随机游走过程

中导致了误差的传播. 而当已知数据增多的时候,已标注数据分散在未标记数据中,可以比较有效地控制误差的传播,从而减小了分类误差.

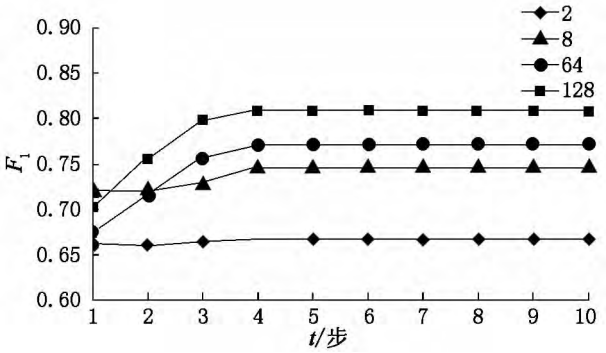

图5 F_1 值随迁移步数的 t 的变化情况

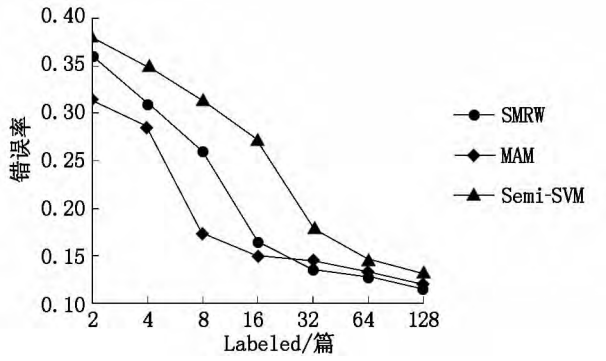

图6 SMRW,MAM 和 Semi-SVM 分类效果

图7 从错误率的角度对比了 SMRW,MAM 和 Semi-SVM 3 种半监督分类方法的分类效果,从图6中可以看出,本文提出的 SMRW 具有较好的分类性能.

图7 从 F_1 值的角度比较了 SMRW 模型和 Semi-SVM 分类性能,从图7可以看出,前者的分类性能要优于后者.

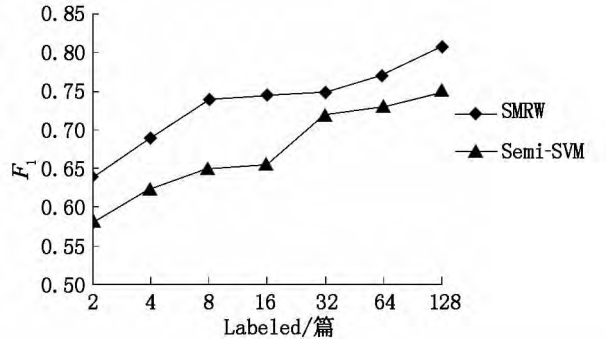

图7 SMRW 和 Semi-SVM 的 F_1 值的结果

图8比较了SMRW模型和加入了渐进学习思想的PSMRW 模型的分类性能,从图8中可以发现,PSMRW 模型较SMRW 模型的 F_1 值都有所提高. 随着标记点数的增加,F_1 值增加的速度首先加快,但当标记点达到一定数量时,速度反而下降,这是因为随着标记点数的增加,半监督引入的误差逐渐减少,渐进学习过程中能修正的错误信息较少. F_1 值在

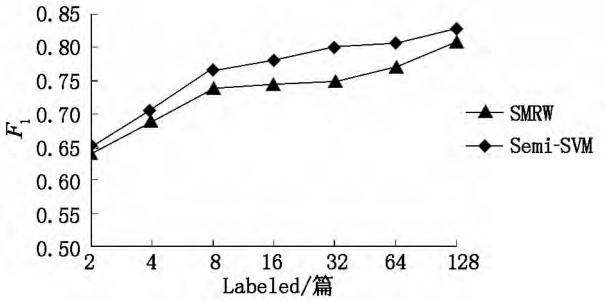

图8 SMRW 和 PSMRW 得到的 F_1 值的结果

$n_l = 32$ 时的增量最多,达到0.05.

4 结束语

本文提出基于 Markov 随机游走半监督分类模型(SMRW),由于迭代过程中,样本被错分所引起的误差,会在之后的迭代中被不断放大,从而造成文本分类的准确性降低. 为了进一步提高模型的分类精度,本文提出了加入渐进学习思想的方法(即PSMRW),该方法通过不断地"纠正"半监督学习过程中的"错误"来改善半监督学习的效果. 实验结果表明,该算法能有效地提高半监督学习的效率. 渐进式半监督分类未来进一步的研究工作主要包括:(ⅰ)尝试将此方法应用在文本的多类分类问题上;(ⅱ)在特征提取时考虑用其他的方法,如对特征加权等方法;(ⅲ)考察用不同的相似度计算方法对其效果的影响.

5 参考文献

[1] Zhu Xiaojin. Semi-supervised learning literature survey [R/OL]. [2013-03-19]. http://www.loni.ucla.edu/~ztu/courses/2013_CS_spring/reading/ssl_survey.pdf.

[2] Zhou Zhihua, Zhan D C, Yang Q. Semi-supervised learning with very few labeled training examples [C/OL]. [2013-03-21]. Semi-supervised learning with very few labeled training Examples.

[3] 易星. 半监督学习若干问题的研究 [D]. 北京: 清华大学,2004.

[4] 董乐红,耿国华,高原. Boosting 算法综述 [J]. 计算机

应用与软件，2006，23(8)：27-29.

[5] Liu Bin，Lee W S，YU P S，et al. Partially supervised classification of text documents [C/OL]. [2013-04-11]. http: // www. cs. uic. edu/ ~liub/S-EM/unlabelled. pdf.

[6] 郑海清，林琛，牛军钰. 一种基于紧密度的半监督文本分类方法 [J]. 中文信息学报，2007，21(3)：54-60.

[7] Bluma，Mitchell T. Combining labeled and unlabeled data with co-training [C/OL]. [2013-04-14]. http: // dl. acm. org/citation. cfm? id = 279962.

[8] Szummer M，Jaakkola T. Partially labeled classification with Markov random walks [J]. Advances in Neural Information Processing Systems，2001，14(1)：1-8.

[9] Arik Azran. The rendezvous algorithm: multi-class semi-supervised learning with Markov random walks [C/OL]. [2013-05-12]. http: // citeseerx. ist. psu. edu/viewdoc/summary? doi = 10. 1. 1. 74. 2110.

[10] Robert E. Schapire，Yoram S. BoosTexter: a boosting-based system for text categorization [J]. Machine Learning，2000，39(2/3)：135-168.

[11] 任巨伟，杨亮，林鸿飞. 情感图式构造及其在文本情感计算中的应用 [J]. 江西师范大学学报: 自然科学版，2013，37(2)：130-135.

[12] 何文译，林鸿飞，杨亮. 基于群体智慧的电影排序模型 [J]. 江西师范大学学报: 自然科学版，2013，37(2)：136-141.

The Progressively Semi-Supervised Classification Model Based on Markov Random Walk

CHEN Xiu-ping[1]，WANG Ming-wen[1*]，WAN Jian-yi[1]，ZUO Jia-li[2]

(1. College of Computer Information Engineering，Jiangxi Normal University，Nanchang Jiangxi 330022，China；
2. School of Elementary Education，Jiangxi Normal University，Nanchang Jiangxi 330027，China)

Abstract: The progressively semi-supervised classification model based on Markov random walk，in the random walk process has been proposed，and calculated the migration probability of samples to be marked，considering only samples of the appropriate category，while ignoring the other classes of samples; and then combined the progressive learning with semi-supervised learning. The model can improve the precision by "correcting" the errors caused in semi-supervised learning process. The results on 20newsgroups dataset in the experiment shows that the proposed method can improve the accuracy of semi-supervised classification.

Key words: semi-supervised classification; progressive learning; Markov random walk; iterating

(责任编辑: 冉小晓)