

一种改进的 ML- k NN 多标记文档分类方法

程圣军, 黄庆成, 刘家锋, 唐降龙

(哈尔滨工业大学 计算机科学与技术学院, 150001 哈尔滨)

摘 要: 针对应用传统 k 近邻算法进行多标记文档分类时忽略了标记之间相关性的问题, 提出了一种改进的 ML- k NN 多标记文档分类方法. 针对文本特征的特点, 采用一种基于 KL 散度的距离尺度来更好地描述文档相似度. 根据近邻样本所属类别的统计信息, 通过一种模糊最大化后验概率法则来推理未标记文档的标记集合. 与 ML- k NN 不同的是, 该方法可以有效地利用标记相关性来提升分类性能. 在 3 个标准数据集上, 5 个多标记学习常用评测指标下的实验结果表明: 所提方法在多标记文档分类问题上要明显优于 ML- k NN、Rank-SVM 和 BoosTexter 等主流多标记学习算法.

关键词: 文档分类; 多标记学习; 标记相关性; k 近邻; KL 散度

中图分类号: TP181

文献标志码: A

文章编号: 0367-6234(2013)11-0045-05

An improved ML- k NN approach for multi-label text categorization

CHENG Shengjun, HUANG Qingcheng, LIU Jiafeng, TANG Xianglong

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

Abstract: Conventional k NN algorithms ignore label correlations when being applied to multi-label text categorization. To cover this shortage, an improved Multi-label k NN approach for text categorization is proposed. A specific distance metric based on KL divergence is derived to measure the similarity between individual documents. Based on statistical information gained from the label sets of neighboring documents, a fuzzy maximum a posteriori principle is utilized to conjecture the label sets of the unlabeled documents. Different from ML- k NN, the proposed approach can exploit label correlations to improve classification performance effectively. Experiments on three benchmark datasets using 5 popular multi-label evaluation metrics suggest that the proposed approach achieves superior performance to some well-established multi-label learning algorithms, such as ML- k NN、Rank-SVM and BoosTexter.

Key words: text categorization; multi-label learning; label correlations; k nearest neighbor; KL divergence

在传统文档分类问题中, 文档具有明确、单一的语义, 监督学习可以有效地解决该分类问题. 然而, 真实世界中文档可能同时具有多个语义, 比如, 一篇新闻稿可能同时拥有多个主题. 多义性对象不仅存在于文档分类中, 还广泛存在于图像场景分类、生物信息学中^[1]. 如何利用具有多义性的数据进行有效学习已成为近年来机器学习研究中的热点之一.

多标记学习^[2]作为一种解决多义性对象学习建模的框架, 其研究成果已经广泛应. 用到文档分类^[3]、web 数据挖掘^[4-5]、生物信息学^[6]等多个领域. 现有的多标记学习方法大致可分为两类^[7]: 1) 问题转换方法. 该类方法的基本思想是通过将多标记训练样本进行处理, 将多标记学习问题转换为其他已知的学习问题进行求解. 文献[2]中介绍了该类几种典型方法; 2) 算法适应方法. 其基本思想是通过将常用监督学习算法进行改进, 将其直接用于多标记数据的学习. 代表性学习算法^[8]有基于神经网络的方法^[9], 基于 Boosting 的 BoosTexter^[3], 基于支持向量机的 Rank-SVM 方法^[10], 基于图的方法^[16-18]和基于 k 近邻的

收稿日期: 2012-05-25.

基金项目: 国家自然科学基金资助项目(61173087, 61073128);
黑龙江省自然科学基金资助项目(F201021).

作者简介: 程圣军(1985—), 男, 博士研究生;
唐降龙(1960—), 男, 教授, 博士生导师.

通信作者: 程圣军, hitwer@gmail.com.

ML- k NN^[12]方法等.

ML- k NN 的主要思想是采用 k 近邻作为分类准则,统计近邻样本的类别标记信息,通过最大化后验概率的方式推理未见样本的标记集合.该方法采用 Euclidean 距离来衡量样本间相似度,但在文档分类问题中,该距离往往不是最优的选择.另外,ML- k NN 忽略了标记间的相关性.而在实际应用中,类别标记之间的相关性^[11,13]是存在的.比如,一篇属于音乐范畴的新闻稿同时很有可能属于娱乐范畴.

针对以上问题,本文提出一种基于 ML- k NN 改进的多标记文档分类方法.采用一种基于 KL 散度的距离描述文档样本相似度.通过模糊理论来改进未见文档标记的推理过程,将文档的类别标记相关性考虑进来.实验结果表明:本文方法优于 BoosTexter、Rank-SVM 和 ML- k NN 等主流多标记学习算法.

1 方法描述

1.1 ML- k NN

首先引入一些相关符号:给定样本 x 及对应标记集合 $Y \subseteq \Omega = \{1, 2, 3, \dots, Q\}$. y_x 为样本 x 的类别向量,即若 l 为 x 的标记($l \in Y$), $y_x(l) = 1$, 反之 $y_x(l) = 0$. $N(x)$ 为 x 的 k 近邻样本所组成的集合. $C_x(l)$ 用于统计样本 x 的近邻中属于第 l 个类别的数量,可表示为

$$C_x(l) = \sum_{a \in N(x)} y_a(l). \quad (1)$$

给定测试样本 t , ML- k NN 首先从训练集中找出它的 k 近邻集合 $N(t)$. 令 H_l^t 表示 t 有标记 l 这一事件,而 H_0^t 表示 t 没有标记 l 这一事件, $E_j^l (j \in \{0, 1, \dots, k\})$ 表示 t 的 k 近邻中恰有 j 个样本具有标记 l 这一事件. y_t 可由 MAP 准则得到

$$y_t(l) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l | E_{C_t(l)}^l), l \in \Omega. \quad (2)$$

对式(2)使用贝叶斯法则可得

$$y_t(l) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l) P(E_{C_t(l)}^l | H_b^l). \quad (3)$$

根据式(3),为确定测试样本 t 的类别信息,只需根据训练集中近邻类别的统计频数来计算先验概率 $P(H_b^l)$ 以及后验概率 $P(E_{C_t(l)}^l | H_b^l)$.

1.2 基于标记相关性改进的 ML- k NN

ML- k NN 的潜在假设为:标记之间互相独立,不存在相关性.该算法通过式(3)判断测试样本 t 是否属于类别 l 时,仅考虑近邻中属于类别 l 的情况,即条件概率 $P(E_{C_t(l)}^l | H_b^l)$,完全忽略了近邻中其他类别的统计信息.为了考虑标记间的相关性,本文通过一种模糊最大化后验概率(Fuzzy MAP)

准则来重写 $y_t(l)$,可得

$$y_t(l) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l | \{E_{C_t(q)}^q\}_{q \in \Omega}) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l | E_{C_t(l)}^l, \{E_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}}). \quad (4)$$

式中 $E_j^q (j \in \{0, 1, \dots, k\})$ 表示 t 的 k 个近邻中恰有 j 个样本具有标记 q 这一事件.从式(4)中可以看出,样本 t 是否属于 l 类别不仅取决于事件 $E_{C_t(l)}^l$,而且取决于 $\{E_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}}$. 以此将标记相关性考虑进来.

由于无法直接计算 $\{E_{C_t(q)}^q\}_{q \in \Omega}$,本文采用模糊理论来近似计算式(4).首先定义模糊事件 F_j^q , $j \in \{0, 1, \dots, k\}$,即 t 近邻中有近似 j 个具有标记 q 的近邻的事件, $C_t(q)$ 位于区间 $[j - \delta_q; j + \delta_q]$ 内,其中 $\delta_q \in \{0, \dots, k\}$. 通过 F_j^q 近似描述事件 E_j^q ,式(4)可重写为

$$y_t(l) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l | E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}}). \quad (5)$$

文中标记间相关性是通过向量 $\delta = (\delta_1, \dots, \delta_Q)$ 来描述的, δ_q 越小,标记 l 与 q 之间的相关性越大.通过贝叶斯法则改写式(5),可得

$$y_t(l) = \operatorname{argmax}_{b \in \{0, 1\}} P(H_b^l) P(E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}} | H_b^l). \quad (6)$$

另外,为对测试样本 t 的标记进行排序,需计算其对应各类别标记的后验概率为

$$r_t(l) = P(H_b^l | E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}}) = \frac{P(H_b^l) P(E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}} | H_b^l)}{\sum_{b \in \{0, 1\}} P(H_b^l) P(E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}} | H_b^l)}. \quad (7)$$

对比式(3)和式(6),改进算法与原始 ML- k NN 的差别在于后验概率的计算,改进算法通过修改后验概率公式以达到引入标记相关性的目的.根据式(6)可确定测试样本 t 是否具有 l 标记,式中先验概率和后验概率可根据训练集中样本的类别统计信息而得,其中计算后验概率 $P(E_{C_t(l)}^l, \{F_{C_t(q)}^q\}_{q \in \Omega \setminus \{l\}} | H_b^l)$ 时,仅考虑满足以下条件的训练样本 x 为

$$\begin{cases} C_x(l) = C_t(l), \\ C_t(q) - \delta_q \leq C_x(q) \leq C_t(q) + \delta_q, q \in \Omega \setminus \{l\}. \end{cases} \quad (8)$$

$\delta = (\delta_1, \dots, \delta_Q)$ 可以通过在训练集上运行交叉验证法(cross validation)得到.

图1为本文算法步骤, s 为平滑系数,根据文献[12]中设置,本文 s 取值为1.

近邻 $N(x)$ 的选择很大程度上决定了算法性能. ML- k NN 通过 Euclidean 距离来刻画样本间相

似度,而事实上,该距离并不能很好地描述文档间的相似度.

输入: 训练集 T , 近邻数 k , 测试样本 $t, s, \delta = (\delta_1, \dots, \delta_Q)$
输出: $y_t(l), r_t(l)$
初始化: For $q = 1, \dots, Q$
 计算 T 中所有类别的先验概率 $P(H_0^q), P(H_1^q)$
 通过式(1) 计算 $C[q], C'[q] = |T| - C[q]$
 EndFor
1. 对每个测试样本 t , 定位其近邻 $N(t)$, 统计 C_t
2. 对所有 $x \in T$, 计算 $N(x), C_x$
3. For $l = 1, \dots, Q$
4. 统计 $C_x[l]$ 满足式(8) 的训练集 T 中的样本数目 $V[l]$
5. 统计不满足式(8) 的训练集 T 中的样本数目 $V'[l]$
6. $P(E_{C_t(l)}^l, \{F_{C_x(q)}^y\}_{q \in \Omega \setminus \{l\}} | H_1^l) = (s + V[l]) / (s \times Q + C[l])$
7. $P(E_{C_t(l)}^l, \{F_{C_x(q)}^y\}_{q \in \Omega \setminus \{l\}} | H_0^l) = (s + V'[l]) / (s \times Q + C'[l])$
8. EndFor
9. For $l = 1, \dots, Q$
10. 分别根据式(6) 计算 $y_t(l)$, 式(7) 计算 $r_t(l)$
11. EndFor

图 1 算法的具体步骤

1.3 基于 KL 散度的文档相似度描述

在文档分类问题中,文本样本一般采用 bag of words 作为特征表示形式,即将文档转化为基于词频统计的特征向量^[14]. 给定一文档 d_i, w_t 为词汇表 V 中第 t 个词,记 $N(w_t, d_i)$ 为词 w_t 在 d_i 中出现的次数,则 $N(w_t, d_i)$ 为文档 d_i 的特征向量中第 t 个分量.一般描述文档相似度的方法为计算特征向量间的距离 (Euclidean、向量夹角余弦). 本文采用指数形式的 KL 散度来描述文档的相似度. 两文档 d_i 和 d_j 的相似度为

$$\text{sim}(d_i, d_j) = e^{-BD(P(w|d_j) \parallel (\lambda P(w|d_i) + (1-\lambda)P(w)))}.$$

(9)

式中: $w \in V; P(w|d_i)$ 为词在文档 d_i 上的最大似然概率分布 (如, $P(w_t|d_i) = N(w_t, d_i) / |d_i|$); $P(w)$ 为词在整个样本集的边缘概率分布; λ 为平滑系数 (避免距离无穷大); β 为控制系数; $D(\cdot \parallel \cdot)$ 是信息论中的一种衡量两个分布之间差异的尺度. 两个分布 $P_1(C)$ 和 $P_2(C)$ 之间的 KL 散度为

$$D(P_1(C) \parallel P_2(C)) = \sum_{j=1}^{|C|} P_1(C) \log \left(\frac{P_1(c_j)}{P_2(c_j)} \right).$$

该方法实质上是通过衡量两文档的“重叠 (overlap)”程度来描述它们之间的相似度;重叠程度越高,相似度越大.本文通过基于 KL 散度的距离尺度来衡量文档相似度,然后利用基于模糊理论改进的 ML- k NN 推理未标记文档的类别标记.

2 实验与分析

本文在 3 个“yahoo.com”数据集上进行实验, 这些数据集为多标记文档分类领域中的 benchmark.数据集的内容来自于真实的网页文档,3 个数据集分别对应 Business&Economy, Science 和 Social&Science 3 个一级类别,每个数据集中的网页又同时属于多个二级类别,因此每个网页都可能有多多个标记.每个数据集都包括 2 000 个训练样本和 3 000 个测试样本,其中约 20%~43% 的文档是多标记的.为了特征降维,仅取文档频率 (document frequency) 最高的前 2% 的词构成最终词汇表.数据集的具体描述如表 1 所示.

表 1 Yahoo 数据集描述

数据集	类别	属性	多标记 文档/%	文档的平均 标记数目/%
Business&Economy	30	438	42. 20	1. 590
Science	40	743	34. 85	1. 489
Social&Scienc	39	1 047	20. 95	1. 274

实验中,式(9)中参数值的设定并不影响样本近邻的定位,因此按照文献[15]的设定, λ 取 0.5, β 取 3.

首先,为了分析不同的近邻数 k 以及不同的相似度刻画尺度对本文方法性能的影响,利用两种不同距离尺度来构造对比算法:“对比 1”(Euclidean 距离),“对比 2”(向量夹角余弦).特征向量夹角余弦值常被用来描述两文档的相似度.需要说明的是,这两种对比算法与本文方法的区别仅在于文档相似度刻画尺度不同.图 2 给出了在不同的 k 取值下算法在 Business&Economy 数据集上的平均精度.其中 δ 可通过在训练集上进行 3 倍交叉验证来确定.

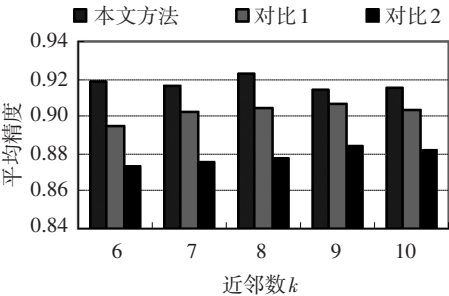


图 2 算法在不同 k 取值下的平均精度

由图 2 所示实验结果可知:
1) 基于 KL 散度的文档相似度刻画优于

Euclidean 距离和向量夹角余弦;

2) 近邻 k 的取值分别为 $\{6, 7, 8, 9, 10\}$ 时, 各算法的平均精度 (average precision) 波动不大, 可见 k 取值对算法性能没有显著影响. 因此, 在本文实验中, 近邻数 k 设定为 8.

本文选择 ML- k NN、Rank-SVM、BoosTexter 作

为对比算法, 采用多标记学习中常用的 5 个评测指标 (Hamming Loss、One-Error、Coverage、Ranking Loss、Average Precision) 来进行比较. 该 5 种评测指标的具体形式详见文献 [3]、[12]. 为了进行统计显著性检验, 各算法分别在 3 个数据集上运行 10 次 (10 次随机选取训练集).

表 2 本文方法与 ML- k NN 的性能比较 (均值±标准差)

评测指标	数据集					
	Business&Economy		Science		Social&Science	
	本文算法	ML- k NN	本文算法	ML- k NN	本文算法	ML- k NN
Hamming loss ↓	0.027 ±0.000	0.031±0.000	0.032 ±0.000	0.034±0.001 0	0.021 ±0.002	0.024±0.001
One-Error ↓	0.110 ±0.001	0.131±0.024	0.574 ±0.015	0.581±0.022 0	0.325 ±0.013	0.337±0.017
Coverage ↓	2.175 ±0.054	2.245±0.031	6.034 ±0.106	6.047±0.018 8	2.828 ±0.112	3.034±0.245
Ranking loss ↓	0.035 ±0.002	0.038±0.007	0.116 ±0.003	0.127±0.009 0	0.059±0.001	0.056 ±0.001
Average Precision ↑	0.923 ±0.002	0.880±0.015	0.533 ±0.026	0.538±0.038 0	0.764 ±0.008	0.728±0.022

表 3 本文方法与 Rank-SVM 的性能比较 (均值±标准差)

评测指标	数据集					
	Business&Economy		Science		Social&Science	
	本文算法	Rank-SVM	本文算法	Rank-SVM	本文算法	Rank-SVM
Hamming loss ↓	0.027 ±0.000	0.0293±0.011	0.032 ±0.000	0.038±0.005	0.021 ±0.002	0.0243±0.002
One-Error ↓	0.110 ±0.001	0.131±0.002	0.579±0.015	0.521 ±0.012	0.325±0.013	0.314 ±0.009
Coverage ↓	2.175 ±0.054	2.415±0.021	6.034 ±0.106	6.691±0.111	2.828 ±0.112	3.687±0.238
Ranking loss ↓	0.035 ±0.002	0.062±0.014	0.116 ±0.003	0.131±0.001	0.059 ±0.001	0.068±0.003
Average Precision ↑	0.923 ±0.002	0.864±0.000	0.538 ±0.026	0.501±0.014	0.764 ±0.008	0.726±0.021

表 4 本文方法与 BoosTexter 的性能比较 (均值±标准差)

评测指标	数据集					
	Business&Economy		Science		Social&Science	
	本文算法	BoosTexter	本文算法	BoosTexter	本文算法	BoosTexter
Hamming loss ↓	0.027±0.000	0.021 ±0.011	0.032 ±0.000	0.034±0.015	0.021 ±0.002	0.024±0.000
One-Error ↓	0.110 ±0.001	0.121±0.015	0.579 ±0.015	0.581±0.022	0.325±0.013	0.312 ±0.007
Coverage ↓	2.175±0.054	2.145 ±0.024	6.034±0.106	6.027 ±0.018	2.828 ±0.112	3.034±0.042
Ranking loss ↓	0.035±0.002	0.028 ±0.000	0.116±0.003	0.127±0.004	0.059±0.001	0.046 ±0.009
Average Precision ↑	0.923 ±0.002	0.880±0.013	0.538±0.026	0.737 ±0.028	0.764 ±0.008	0.728±0.023

表 2、3、4 分别给出了本文算法与 3 种对比算法结果比较. 其中, “↓”表示值越小, 性能越高; “↑”表示值越大, 性能越高; “粗体”表示最优结果. 需要注意的是, 这些数据集中的标记之间并不相互独立. 显著程度为 95% 的 t -检验表明, 在 3 个数据集上, 本文算法在绝大多数指标上都显著优于 ML- k NN 和 Rank-SVM 算法; 与 BoosTexter 的表现相当. 实验结果表明: 在 ML- k NN 推理未见文档的标记时, 通过考虑标记相关性可以取得更优的分类表现.

3 结 论

1) 针对多标记文档分类问题的特点, 本文在传统 k 近邻算法基础上, 提出了一种改进的多标记 ML- k NN 方法. 该方法采用基于 KL 散度的距离来描述文档之间相似度, 通过一种模糊最大化后验概率法则来推理未见文档的标记集合.

2) 实验结果表明, 利用基于 KL 散度的距离来描述文档相似度的效果要明显优于 Euclidean 距离和向量夹角余弦距离.

3) 在 3 个不同数据集上, 通过 5 个评测指标来对比分析算法性能。结果表明: 本文方法要明显优于 ML- k NN, Rank-SVM 和 BoosTexter 等主流多标记学习算法。

参考文献

- [1] TSOUMAKAS G, KATAKIS I. Multi-label classification: an over-view [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.
- [2] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Mining Multi-label Data [M]. Data Mining and Knowledge Discovery Handbook. Berlin: Springer, 2010, 667-686.
- [3] SCHAPIRE R E, SINGER Y. BoosTexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(23): 135-168.
- [4] OZONAT K, YOUNG D. Towards a universal marketplace over the web: Statistical multi-label classification of service provider forms with simulated annealing [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, 2009, 1295-1304.
- [5] HÜLLERMEIER E, FÜRNKRANZ J, CHENG Weiwei, *et al.* Label ranking by learning pairwise preferences [J]. Artificial Intelligence, 2008, 172 (16/17): 1897-1916.
- [6] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function [J]. Bioinformatics, 2006, 22(7): 830-836.
- [7] READ J, HOLMES G, FRANK E. Classifier chains for multi-label classification [J]. Machine Learning, 2011, 85(3): 333-359.
- [8] HUANG Shengjun, ZHOU Zhihua. Multi-label learning by exploiting label correlations locally [C]// Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI' 12). Toronto, Canada: AAAI, 2012: 949-955.
- [9] ZHANG Minling, ZHOU Zhihua. Multi-label neural networks with application to functional genomics and text categorization [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351.
- [10] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification [C]// Advances in Neural Information Processing Systems 14 (NIPS' 01). Cambridge, MA: MIT Press, 2002: 681-687.
- [11] 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, NY: ACM, 2010: 999-1008.
- [12] ZHANG Minling, ZHOU Zhihua. ML- k NN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.
- [13] MA Haiping, CHEN Enhong, XU Linli, *et al.* Capturing correlations of multiple labels: a generative probabilistic model for multi-label text data [J]. Neurocomputing, 2012, 92: 116-123.
- [14] 姜远, 周志华. 基于词频分类器集成的文本分类方法 [J]. 计算机研究与发展, 2006, 43(10): 1681-1687.
- [15] MCCALLUM A, NIGAM K. Employing EM in pool-based active learning for text classification [C]// Proceedings of the International Conference on Machine Learning. San Francisco, CA: Morgan Kaufman Publishers Inc., 1998: 350-358.
- [16] QI Guojun, HUA Xiansheng, RUI Yong, *et al.* Correlative multi-label video annotation [C]// Proceedings of the 15th ACM International Conference on Multimedia. New York, NY: ACM, 2007: 17-26.
- [17] 郑伟, 王朝坤, 刘璋, 等. 一种基于随机游走模型的多标签分类算法 [J]. 计算机学报, 2010, 33(8): 1418-1426.
- [18] KONG Xiangnan, Ng MICHAEL K, ZHOU Zhihua. Transductive multi-label learning via label set propagation [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 704-719.

(编辑 张 红)