相关信息加权的自适应多标签分类算法

周 浩 李 翔 刘功申
(上海交通大学信息安全工程学院 上海 200240)

摘 要 在文本分类中,传统单标签分类问题的解决方法无法简单地应用于多标签文本分类,现有的方法通常会通过单标签问题转化思想或者多标签自身算法改进实现对多标签的文本分类。提出一种相关信息加权的自适应多标签分类算法,该算法具有相关信息加权、自适应阈值调整、权重投票相结合的特点。实验结果表明,该算法的某些性能指标优于现有一些常用的多标签分类方法。

关键词 多标签分类 特征选择 自适应回归 相关信息加权投票

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.01.060

ADAPTIVE ALGORITHM FOR MULTI-LABEL CLASSIFICATION BASED ON RELATED INFORMATION WEIGHTING

Zhou Hao Li Xiang Liu Gongshen
(*School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

Abstract In text classification area, the solution of traditional methods for single-label classification cannot be simply applied to multi-label text classification, and current methods usually implement the text classification of multi-label by improving the single-label problem transformation idea or the multi-label algorithm its own. What is put forward in this paper is an adaptive multi-label classification algorithm with related information weighting, it features in the combination of related information weighting, adaptive threshold adjustment and voting by weight. Experimental results show that some performance indexes of the algorithm are superior to current common multi-label classification methods.

Keywords Multi-label classification Feature selection Adaptive regression Related information weighted voting

0 引 言

目前,单标签分类已经无法满足日益增长的海量多标签数据分类的需求,传统的文本分类方法[1]无法简单有效地应用于多标签分类中,尤其是对中文文本的分类,已逐渐成为受到广泛关注的研究热点。在实际应用中,对日常新闻,各类文章的多标签分类更能反映文本的全面特性[2]。

文本训练向量[3]的表示方式为 $X_i=(x_1,x_2,\cdots,x_n)$, $x_i\in R^n$,其对应的标签集表示为 $Y_i=(y_1,y_2,\cdots,y_m)$, $y_i\in\{0,1\}$。当样本属于第 J 类时, $y_j=1$,不属于第 j 类时, $y_j=0$。单标签分类问题即为多标签分类的一个当 Y 向量的值中只有一个 1 时的特例。

多标签分类[4]指的是,由输入训练数据集定义相关多标签分类器后得到机器学习预测的标签集,使其与实际标签集更为接近。多标签自适应阈值调整[5]是指:根据多标签阈值结果设定的测试迭代,当输入未分类样例数据 $D_e\in X$ 时,对于任意的 $Y_i\in Y$,获得置信系数 $g(x,y)$,多次线性去随机化后得到置信系数使其总体结果与真实情况最为接近。目前,有已被广泛认可的衡量分类结果的正确性与精确度多标签测试指标,例如 Hamming Loss、One-Error、Ranking Loss、Coverage、Average Precision 等等[6]。

针对多标签分类大致有两大主要策略[7],基于问题转化的方法和基于算法转化的方法。前者是将一个多标签问题转化成一组单标签问题后运用已有的单标签分类方法解决,其最大的优势在于灵活性,通过从现有的单标签分类器直接抽象成一个特定的分类器来适应需求。常见的有 BR(BinaryRelevance)、基于标签对比 PW(pairwise comparison)、LP(LabelPowerset)等算法。BR 算法的优势在于概念上的简单和相对快速,但却被认为其脱离了标签间的相关信息,PW 算法的缺点在于其时间复杂度过大,LP 算法的缺点在于其只能对新例子进行分类,而对训练集中的例子过度拟合。后者则是通过改变已有的单标签分类算法,从而使其能够处理多标签数据,如 AdaBoost. MH 算法,其对由简单决策树算法产生的弱规则进行加强,经若干次迭代后,得到一个准确度更高的规则,但训练速度慢,难以处理大文本量信息、ML-kNN 算法、贝叶斯算法等等,它们训练速度快,但若原始语料出现较大的类别偏差,会降低效率[8]。

本文结合了问题转化和多标签算法改进的思想,提出的是一种在各类特征选择基准调整后,基于已有单标签分类结果进行加权、自适应阈值设定,不同权重投票相结合的方法,对待分类实例进行多标签分类,能提高多标签文本分类的准确度与

收稿日期:2013 - 05 - 11。国家自然科学基金项目(61272441, 61171173)。**周浩**,硕士生,主研领域:网络舆情,文本处理。**李翔**,副教授。**刘功申**,副教授。

精度。

1 文本分类的工作基础

1.1 汉语文本自动分词

本文采用的是最大正向匹配的中文分词算法[9]，相当于分词粒度等于零。若在分词词典中的最长词有 k 个汉字字符，将用被处理文本的目标字符串中的前 i 个字作为匹配字段查找字典。若字典中存在这样的一个 K 字词，即为匹配成功，作为一个词切分出来。如果词典中找不到这样的一个 k 字词，即为匹配失败，将匹配字段中的最后一个字去掉，对剩下的字符串重新进行匹配处理……如此迭代进行下去，直到匹配成功，切分出一个词或剩余字串的长度为零为止。然后取下一个 K 字字符串进行匹配处理，直到文本扫描完毕。

1.2 特征选择

首先对文本粗降维，指的是训练文本经分词后首先去掉停用词，即一些没有实际分类意义的高频词、稀有词。高频词会多次出现在各种类别的文本中，稀有词属于偶尔出现在各个类别中，没有实际的分类检索意义，同时清除些多余的符号等冗余。本文中采用了经过各类别字词贝叶斯统计分析后，建立停用词表，通过词表法去掉高频词和稀有词[10]。

但是文本的向量空间表示初始维数依旧有可能太大，会导致维度爆炸，我们必须对向量空间进行降维。特征选择的目的就是从原有的特征中选择出与标签集有最大的依赖度的子集。根据 TF-IDF 公式[11]计算互信息量的方法选择最具有类别识别度的特征，从而选出对分类贡献重大的特征向量，与此同时也可以提高分类的精度和程序的运行速率及效率。

$$W(t,d) = \frac{(1+\log_2 tf(t,d)) \times \log_2(N/n_i)}{\sqrt{\sum_{t\in d}[(1+\log_2 tf(t,d)) \times \log_2(N/n_i))]^2}} \quad (1)$$

其中，$w(t,d)$ 为词 t 在文本 d 中的权重，$tf(t,d)$ 为词 t 在文本 d 中的词频，N 为训练文本集的训练文本总数。n_i 为训练文本集中出词 t 的文本数。分母为归一化因子。我们需要对向量空间进行降维，保留那些对分类贡献重大的词，提高分类的精度。同时也可以提高程序的运行速率和效率。

1.3 特征权重调整

为了更好地选取文本特征，我们必须尽可能选出在各个类别中具有代表性的词，为了达到其目的，采用了一种特征权重调整的策略，为每个前期特征选择后的特征赋予权值，设定为该特征在最大类别中的出现频率与所有类别出现频率的平均值的比值，$WC_i = \frac{Maxf(C_i)}{\arg\sum_{i=1}^{|C|} f(C_i)}$，使得在文本中更能体现类别特点的词获得更大的权重。

2 相关信息加权的自适应多标签分类算法

2.1 信息加权模型算法

信息相关模型加权的基本思想是从一个文本出发，随机找到其相邻文本，并计算出文本间的距离作为权重。遍历其在一定距离范围内的邻居文本，反复迭代后得到一个与初始文本相关度最大的各个文本并得到距离概率分布。

首先将训练集合 D 映射成模型图中的一个点集合 V，对于待处理点计算其 v_i 相邻点的欧氏距离并且将其相连，基于欧氏距离的相似概率可定义为：

$$SIM_E(d_u,d_i) = \sqrt{\sum_{j=1}^{|C|} |c_{uj}-c_{ij}|^2} \quad (2)$$

模型图可表示为：图 G 中有点集合 V，其包含的边为距离在一定范围内的相邻点。

$$\begin{aligned} &G = (V,E) \\ &V = \{v_i \mid x_i \in X, 1 \leqslant i \leqslant m\} \\ &E = \{(v_i,v_j) \mid v_i,v_j \in V, Y_i \cap Y_j \neq \Phi, i \neq j\} \end{aligned} \quad (3)$$

其权重值表示为 W_{ij}：

$$W_{ij} = \begin{cases} 0 & v_i = v_j \\ \infty & v_i \neq v_j \quad (v_i,v_j) \notin E \\ dis(v_i,v_j) & v_i \neq v_j \quad (v_i,v_j) \in E \end{cases} \quad (4)$$

例如：根据一个四类标签集合语料，$Y = \{y_1,y_2,y_3,y_4\}$，训练数据集中包含四个文本实例，文本类标签为表 1 所示，模型计算的新加入实例与各个类标签间的权重为表 2 所示。

表 1 训练集合实例

实例序号	标签集合
1	$Y_1 = \{y_1,y_2,,y_4\}$
2	$Y_2 = \{y_1,y_4\}$
3	$Y_3 = \{y_2,y_3\}$
4	$Y_4 = \{y_2,y_3,y_4\}$

表 2 类标签间的权重

测试实例	1	2	3	4
S	0.95	0.84	0.73	0.68

任意两个样例在特征集合表现出相似性时，那么它们在类标签集合上也会具有相似性。由于多类标数据集的类标维度大于 1，有时甚至和特征集合的维度相当，上面的特性反过来也成立：就是说任意两个样例在类标集合上具有相似性时，在特征集合也会表现出相似性。因此根据这个特性提出加权属性调节权值的方法：分别对训练数据集特征空间的每个特征分量进行分析，计算每个样例在缺少这个特征分量时的多个近邻，得到的类标签集合与这个样例基于类标的多个近邻类标集合。

2.2 WeightedLabelPower 投票预测

这个方法即将多标签问题转换为单标签多类分类问题，转换类的属性值与训练样本实例的标签集相关，基于投票机制，对所属文本进行类标签判断，如表 3 所示，总计大于阈值 $K \times 0.5$ 的即为预测标签，此例为 $K=4$。

表 3 WeightedLabelPower 投票预测

实例序号	Y_1	Y_2	Y_3	Y_4
1	0.95	0.95	0	0.95
2	0.84	0	0	0.84
3	0	0.73	0.73	0
4	0	0.68	0.68	0.68
总计	1.79	2.35	1.41	2.47
标签预测	0	1	0	1

2.3 多标签分类算法的框架描述

对于训练集的样本特征进行统计后得到每个特征的权重调

整,从而使特征更能反应其类别特性。为每个测试实例通过调整后的权重特征,找到其在训练集中相应的 K 个邻居实例,将它们与其 K 个邻居节点间的距离作为类别实例权重。通过 WeightedLabelPower 投票策略,预测出分类结果。对于总体结果进行统计性能测试,基于 Hamming Loss、Ranking Loss、Coverage、Average Precision、One-Error 的总体评价,调整邻居节点的数目,反复迭代得出结果。

Algorithm:IWLC

Input: D:Multi-label trainset、S:test example set

output:Y:predicted label set for S

Process:

Step1 //计算每个权重的特征调整

For $i=0$ to n

Computer the WCi for feature

End for

Step2 //计算并选择根据欧氏距离找出样本 s 的邻居节点

$N(d_u) \Leftarrow \Phi$

For $j=0$ to k

$n_i \Leftarrow \underset{dj \in \Omega - N(du)}{\operatorname{argmin}} SIM(d_u, d_j)$

//标准化 S_k 与样本间各距离差值

$$D(s,d_i) = \sqrt{\sum_{i=1}^{n} (Sc_i - Xc_i)^2}$$

$$D_j = D(s,d_i)/\max D(s,d_i)$$

End for

Step3 //找出 K 个距离最小的实例和其标签集作为点的邻居图,将//距离值作为其权重值

For $i=0$ to $|S|, j=0$ to k

Get the $S_{ij}[\,][\,]$ who holds the neighbour and weight

End for

Step4 //按照 WeightedLabelPower 投票机制找出初步预测的标签集

For $i=0$ to $|S|, j=0$ to k

Get the $T[i][j]$ who holds the preliminary forecasting

End for

Step5 //对于预测出的各个样本标签集做各方法的性能测试

For $i=0$ to $|S|, j=0$ to k

Ranking[i][j],Confidence[i][j],Truelabels[i][j]

End for

Test the Hamming Loss、Ranking Loss、Coverage、Average Precision、One-Error

Step6 //反复根据 step4 的结果做更改每个类的 K 值的 Step2—Stpe4 的迭代,找出最佳方案

3　实验结果及分析

3.1　多标签性能测量指标

本文选取的多标签性能指标为 Hamming Loss、One-Error、Ranking Loss、Coverage、Average Precision,如表 4 所示。Hamming-gloss 指的是实例真实结果与实例预测结果集间的异或,此评价代表了实例标签对错分类的次数;One-error 是指该预测实例类别相关度最高的类与实际结果的异或,此评价代表了最高排名的标签不在例子实际分类中的次数;Coverage 指的是正确结果的错误度,此评估代表了平均每个预测实例需要降低多少格才能找到精确的标签;Ranki-loss 指的是评估了平均标签对的局部排序错误,该评估反应了预测结果在排名上的错误,Average-percision 评估了预测出的标签平均精确程度。前四个方面评估值越小越好,但最后的 Average-percision 值是越大表现越好。

表 4　评价指标

名称	公式
Hamming-loss ↓	$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{Q}\lvert h(x_i)\Delta Y_i\rvert$ 其中 $h(X_i)\Delta Y_i$ 分类结果为分类结果和实际样本的差集
Ranking-loss ↓	$\frac{1}{m}\sum_{i=1}^{m}\frac{\lvert R(x_i)\rvert}{\lvert Y_i\rvert\lvert\overline{Y_i}\rvert}$ $R(x_i)=\{(l_0,l_1)\mid\frac{f(x_i,l_1)}{f(x_i,l_0)}\leqslant 1$ $(l_0,l_1)\in Y_i\times\overline{Y_i}\}$
Coverage ↓	$\frac{1}{m}\sum_{1}^{m}\lvert C(x_i)-1\rvert$ $C(x_i)=\{1\mid f(x_i,1)\geqslant f(x_i,l_i'),l\in y\}$ $l_i'=\operatorname{argmin}_{k\in Y_i} f(x_i,k)$
One-error ↓	$\frac{1}{m}\sum_{i=1}^{m}H(x_i)$ 结果中第一名在实际标签中 $H(x_i)=1$
Average Percision ↑	$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{Y_i}P_i$ $P(x_i)=\sum_{k\in Y_i}\frac{\lvert\{l\mid f(x_i,l)\}\geqslant f(x_i,k),l\in Y_i\}\rvert}{\lvert\{l\mid f(x_i,l)\}\geqslant f(x_i,k),l\in y\}\rvert}$

注:↑表示值越大效果越好,↓表示值越小效果越好。

3.2　语料描述

本文采用的是酵母[12]、景象[13]和情感[14]英文数据集和一个来自同济大学卫志华老师提供的中文新闻文本语料库[15],其具体的信息包括训练样本数、测试样本数、样本特征数、标签数、及平均标签长度,如表 5 所示。酵母数据集是一个关于基因功能分类的数据集,其中每个样本代表一个基因,它的特征来自于基因的微阵列表示和系统发育谱;景象数据集的每个样本代表一幅图像,样本的特征取 自于图像的颜色信息和结构信息;情感数据集的每个样本代表人们听到某种音乐所产生的情感,样本的特征取自于音乐的节奏和音色。中文文本语料库的样本是取自教育,经济,军事,科技,商务,社会,体育,娱乐,政治共九大类的中文文本新闻数据集。现实的新闻语料的多标签情况受到许多因素的影响,如在人工划分对内容理解的主观影响、概念区分不清晰、标签之间从属关系等。数据本身就存在大量噪声。此外,在多标签数据中各类样本分布很不均匀,所以要尽量选取较为平均分布的语料。

表 5　数据实例集描述

数据集	Emotions	Scene	Yeast	同济新闻语料库
所属领域	音乐	图像	生物	新闻
文档数	593	2407	2417	5894
特征数目	72	294	103	2344
标签数	6	14	14	9
标签势	1.869	1.074	4.237	1.197
标签密度	0.311	0.179	0.303	0.199
不同标签集数	27	15	198	125

表 5 中的标签势是指训练数据集中实例的平均标签数目,而标签密度指的是标签势数除以标签数。

$$LC(D) = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\sum_{j=1}^{|L|} l_{ij} \quad (5)$$

$$LD(D) = \frac{1}{|\Omega||L|}\sum_{i=1}^{|\Omega|}\sum_{j=1}^{|L|} l_{ij} \quad (6)$$

3.3 实验结果

本文实验环境为 Intel(R)Xero CPU E5620@ 2.40 GHz,15.9 GB内存,1T 硬盘的华为服务器,操作系统为 Winserver 2003,Java 版本 Sun JDK 1.7.0。采用 10 倍交叉验证(10—fold Cross—validation)策略对四个数据集进行了仿真实验。根据自适应迭代测试,情感、景象和酵母数据集和同济新闻语料库的初始 K 值分别选定为 10、10、10、15。实验中与 IWLC 算法采用的对比算法有 MLkNN[16](Multi-Label k-nearest neighbor)、BRkNN[6](Binary Relevance k-nearest neighbor)、RAkEL[17](Random k-Labelsets)、NB[6](Naive Bayes)。在对比实验中,将原有数据集和测试集混合,随机平衡采样各类并排序。

从表 6 到表 9 中的 5 个评价指标中可以看出,在 Emotions 和 Scene 以及 yeast 的大部分指标上 IWLC 均好于 MLkNN、BRkNN、RAkEL、Naive Bayes 方法,说明了在小数据集分类方面 IWLC 有着明显的分类准确性提高;但在 Scene 上的 Hamming-loss 分类效果上略逊于 ML-kNN 和 BR-kNN,这很有可能是因为其语料为图像数据且特征选取代表性不均衡且标签势太小。在非海量数据实例集上测试,IWLC 的 Ranking Loss、Coverage、Average Precision、One-Error 分类效率显著高于其余各种测试方法,说明了其总体实验结果较之于其他方法有排序较好、整体错分率降低、平均精度提高、最高排名分类准确度更高。在对同济大学提供的大信息新闻语料库的测试中,由于数据分布的复杂性和分类算法达到效果的侧重点不同,ML-kNN 和 BR-kNN 在 Hamming-loss 分类效果上稍优于 IWLC,但其他方面,IWLC 方法例如 Rankingloss、One-Error 依旧尤为突出,在 Coverage,Average Percision 上也优于其他方法,故综合比较还是一种较为可行且有效的多标签分类算法。

表 6 Emotions 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.2100 ±0.0175	0.2621 ±0.0183	2.3627 ±0.0832	0.7356 ±0.0289	0.3391 ±0.0630
BR-k NN	0.2049 ±0.0147	0.2494 ±0.0214	2.2919 ±0.1000	0.7497 ±0.0256	0.3188 ±0.0461
RAkEl	0.2555 ±0.0001	0.2589 ±0.0036	2.2125 ±0.0030	0.7083 ±0.0105	0.4469 ±0.0227
Naive Bayes	0.2518 ±0.0010	0.2208 ±0.0128	2.1063 ±0.0322	0.7580 ±0.0195	0.3306 ±0.0377
IWLC	0.1520 ±0.0018	0.0968 ±0.0102	1.4527 ±0.0508	0.8715 ±0.0206	0.1500 ±0.0172

表 7 Scene 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.0946 ±0.0041	0.1535 ±0.0102	0.8674 ±0.0487	0.8044 ±0.0148	0.2833 ±0.0223
BR-k NN	0.0996 ±0.0073	0.1856 ±0.0062	1.0336 ±0.0287	0.7819 ±0.0120	0.3070 ±0.0219
RAkEl	0.2417 ±0.0002	0.2141 ±0.0030	1.1595 ±0.0132	0.6530 ±0.0048	0.5808 ±0.0097
Naive Bayes	0.1601 ±0.0008	0.1468 ±0.0009	0.8214 ±0.0049	0.7710 ±0.0057	0.3781 ±0.0110
IWLC	0.11390 ±0.0015	0.0691 ±0.0021	0.4577 ±0.0019	0.8580 ±0.0139	0.2534 ±0.0114

表 8 yeast 数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.1982 ±0.0020	0.3312 ±0.0108	9.1345 ±0.1949	0.6641 ±0.0065	0.2565 ±0.0049
BR-k NN	0.2077 ±0.0018	0.2977 ±0.0055	8.5229 ±0.0436	0.6831 ±0.0020	0.2681 ±0.0032
RAkEl	0.2857 ±0.0001	0.3278 ±0.0060	8.8445 ±0.1903	0.6093 ±0.0044	0.4390 ±0.0052
Naive Bayes	0.2682 ±0.0002	0.2652 ±0.0027	7.9243 ±0.1171	0.6698 ±0.0003	0.3471 ±0.0135
IWLC	0.1556 ±0.0013	0.1033 ±0.0005	5.7339 ±0,2012	0.8393 ±0.0041	0.1266 ±0.0178

表 9 同济新闻数据集性能比较

性能对比	Hamming-loss ↓	Ranking Loss ↓	Coverage ↓	Average Precision ↑	One-Error ↓
ML-kNN	0.1509 ±0.0047	0.2900 ±0.0067	3.6860 ±0.0884	0.6186 ±0.0133	0.4508 ±0.0297
BR-k NN	0.1605 ±0.0082	0.2596 ±0.0156	3.4267 ±0.1257	0.6502 ±0.0203	0.4141 ±0.0351
RAkEl	0.1712 ±0.0053	0.1612 ±0.0013	2.4136 ±0.0299	0.7049 ±0.0017	0.4440 ±0.0090
Naive Bayes	0.1810 ±0.0107	0.1218 ±0.0097	2.0604 ±0.1132	0.7856 ±0.0127	0.2984 ±0.0181
IWLC	0.1796 ±0.0075	0.1277 ±0.0014	2.3352 ±0.0917	0.8344 ±0.0121	0.1591 ±0.0120

4 结 语

本文采用的一种相关信息加权的自适应多标签分类算法,相对于现有的一些多标签分类方法在大部分性能指标上有所提高。自适应选择的过程会帮助算法在针对不同领域的的语料库有更好的效果,将经典线性回归体系扩展到多标签分类。实验可见,IWLC 算法提供了一种更为有效,分类可靠性更高的多标签分类算法,本文的后续工作是进一步改进其在分类精度上的进一步改善。

参 考 文 献

[1] Godbole S,Sarawagi S. Discriminative methods for muhi-labeled classification [C]//Proceedings of the 8th Pacic-Asia Conference on Knowledge Discovery and Data Mining. 2004,3056:22 -30.

[2] 陈震,吴斌,沈崇玮,等. 一种改进的基于质心的文本分类算法[J]. 计算机应用与软件,2013,30(1):43 -47,54.

[3] 吕小勇. 多标签文本分类算法研究[D]. 山西财经大学,2010.

[4] Streich A,Buhmann J. Classfication of multi-labeled data:A generative approach[C]//Proceedings of the ECMI/PKDD. Antwerp, Belgium, 2008,2:390 -405.

[5] 裴颂文,吴百锋. 动态自适应特征权重的多类文本分类算法研究[J]. 计算机应用研究,2011,28(11):4092 -4096.

[6] Tsoumakas G, Katakis I, Vlahavas I. Multi-Label Classification: An Overview[J]. International Journal of Data Warehousing and Mining, 2007,3(3):1 -13.

[7] Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data. Data Mining and Knowledge Discovery Handbook[M]//Maimon O,Rokach L. Springer,2010:667 -685.

[8] Modi Hiteshri,Panchal Mahesh. Experimental Comparison of Different Problem Transformation Methods for Multi-Label Classification using MEKA[J]. International Journal of Computer Applications ,2012,59(15):10 -15.

[9] 张宁. 基于语义的中文文本预处理研究[D]. 西安电子科技大学,2011.

[10] 符红霞,黄成兵. 采用特征分辨率和等价类相关矩阵的特征选择[J]. 科学技术与工程,2012,12(34):9234 -9237,9242.

[11] Kou H,Gardarin G,Zeitouni K. Approaches to feature selection for document categorization [C]//Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems. Amsterdam, Netherlands: Elsevier Science Publishers. 2003: 141 -154.

[12] Pavlidis P,Weston J,Cai J,et al. Combining mieroarray expression data and phylogenetic profiles to learn functional categoties using support vector machines[C]//Proceedings of Annual International Conference on Computational Molecular Biology. Columbia: Columia University, 2001:242 -248.

[13] Boutell M R,Luo J,Shen X,et al. Learning multi-label scene classification[J]. Pattern Recognition,2003,37(9):1757 -1771.

[14] Trohidis K,Tsoumakas G,Kalliris G,et al. Multi-label classification of music into emotions [C]//Proceedings International Conference on Music Information Retrieva1. Philadelphia:ISMIR,2008:325 -330.

[15] 卫志华. 中文文本多标签分类研究[D]. 上海:同济大学,2010.

[16] Zhang M L,Zhou ZH. A k-nearest neighbor based algorithm for multi-label classification [C]//Proceedings of the IEEE International Conference on Granular Computing. Heidelberg:Springer Berlin,2004:718 -721.

[17] Tsoumakas G,Vlahavas I. Random k-labelsets:an ensemble method for multi-label classification[C]//Proceedings of the 18th European Conference on Machine Learning,2007:406 -417.

(上接第 142 页)

[36] Criminisi A. Accurate Visual Metrology from Single and Multiple Uncalibrated Images[D]. Berlin:Springer-Verlag,2001.

[37] 张祖勋. 数字摄影量测与计算机视觉[J]. 武汉大学学报信息科学版, 2004,29(12):1035 -1039.

[38] 王美珍,刘学军,甄艳,等. 基于交比的单幅图像平面几何信息提取算法研究[J]. 武汉大学学报:信息科学版,2011,36(2):190 -194.

[39] Toyama K,Lgoan R,Roseway A. Geographic location tags on digital images[C]//Proceedings of the Eleventh ACM International Conference on Multimedia,2003:156 -166.

[40] Naaman M,Song Y J,Paepcke A, et al. Automatic organization for digital photographs with geographic coordinates[C]//Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries,2004,6:53 -62.

[41] Yekkala A K,Volleberg G T G,Saha S. Automatic organization of digital photographs[J]. Consumer Electronics,2007:1 -2.

[42] Epshtein B,Ofek E,Wexler Y,et al. Hierarchical photo organization using geo-relevance[C]//Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information systems, 2007:1 -7.

[43] Pongnumkul S,Wang J,Cohen M. Creating map-based storyboards for browsing tour videos[C]//Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology,2008: 13 -22.

[44] Paul Lewis. Linking spatial video and GIS[D]. Maynooth,Ireland: National University of Ireland Maynooth,2009.

[45] Wu Yong,Liu Xuejun,Lin Guangfa. The Locatable Video:Acquisition, Segmentation, Retrieval [C]//Proceedings of the 19th International Conference on Geoinfomatics,2011:1 -5.

[46] Sakire Arslan Ay,Roger Zimmermann,Seon Ho Kim. Relevance ranking in georeferenced video search[J]. Multimedia Systems Journal, 2010,16(2):105 -125.

[47] Seon Ho Kim,Sakire Arslan AY,Roger Zimmermann. Design and implementation of geo-tagged video search framework[J]. Journal of Visual Communication and Image Representation,2010:773 -786.

[48] 韩志刚. 地理超媒体数据模型及 Web 服务模型研究[D]. 河南:河南大学,2011.

[49] 李德仁,沈欣. 论基于实景视频的城市空间信息服务——以影像城市·武汉为例[J]. 武汉大学学报:信息科学版,2009,34 (2):127 -130.

[50] 孔云峰. 一个公路视频 GIS 的设计与实现[J]. 公路,2007,1(1):118 -121.

[51] Armin Gruen, Zhang Li, Xinhya Wang. 3D city modeling with TLS (Three Line Scanner) data[J]. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2003, Vol. XXXIV-5/W10.

[52] 桂德竹. 基于组合宽角相机低空影像的城市建筑物三维模型构建研究[D]. 北京:中国矿业大学,2010.

[53] Debevec P E,Taylor C J,Malik J. Modeling and rendering architecture from photographs:a hybrid geometry and image-based approach[J]. Computer Graphics,1996:11 -20.

[54] Nister D. Automatic dense reconstruction from uncalibrated video sequences [D]. Stockholm, Sweden: Royal Institute of Technology KTH,2001.

[55] Bougnoux S,Robert L. TotalCalib:a fast and reliable system for off-line calibration of images sequences[C]//Proceeding of Computer Vision and Pattern Recognition,1997.

[56] Cipolla R,Robertson D P,Boyer E G. Photobuilder-3D models of architectural scenes from uncalibrated images[C]//Proceeding of IEEE International Conference on Multimedia Computing and Systems,1999: 25 -31.

[57] Noah Snavely,Steven M Seitz,Richard Szeliski. Modeling the world from internet photo collections[J]. International Journal of Computer Vision,2007:189 -210.