一种基于多模态模型的随机子空间分类集成算法

叶云龙, 杨 明

(南京师范大学 计算机科学与技术学院, 江苏 南京 210097)

[摘要] 分类是当前机器学习的重要研究内容之一, 已取得了一定的进展. 现有的文本分类方法大多基于 VSM 模型, 而 VSM 未能有效地利用隐含在文本中的结构信息. 同时, VSM 下的样本空间常常是高维的, 单一的降维策略可能会丢失有用信息. 为改进现有算法的不足, 提出了一种基于多模态模型的随机子空间分类集成算法 MMRFSEn, 有效地利用文本中的结构信息 (单词分布位置的均值和标准差), 且各基分类器是由随机选择的子空间构建而成. 实验结果表明, 该方法是有效可行的.

[关键词] 多模态, 随机子空间, 分类器集成

[中图分类号] TP 301. 6 [文献标识码] A [文章编号] 1672-1292(2009) 04-0057-06

A Multi-modality-based Random Subspace Classifier Ensemble Algorithm

Ye Yunlong, Yang Ming

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China)

Abstract Text Classification is an important machine learning research, in which some progress has been made. Most of the existing classification methods are based on Vector Space Model (VSM), but VSM does not effectively utilize the structure information hidden in the text samples. At the same time, VSM vectors are often high-dimensional, merely using dimensionality reduction strategy may lead to the loss of the useful information. To overcome the shortcomings of the existing algorithms, we propose an algorithm called Multi-modality-based Random Feature subspace classifier Ensemble (MMRFSEn), which can effectively use the structure information hidden in the text such as the words' s average location and standard deviation, and meanwhile each single classifier is constructed by a randomly selected subspace. The experimental results show that the newly developed method is effective and feasible.

Key words Multi-modality, random subspace, classifier ensemble

随着信息技术的发展, 互联网数据及资源呈现海量特征. 为了有效地管理和利用这些分布的海量信息, 基于内容的信息检索和数据挖掘逐渐成为备受关注的领域. 文本自动分类是信息检索与文本挖掘领域的研究热点与核心技术, 近年来得到了广泛的关注和快速的发展.

20世纪 90年代, 基于机器学习的文本分类方法被广泛运用, 并逐渐成为经典范例[1]. 其中, 有代表性的文本分类方法有 Naive Bayes, 决策树、神经网络等. 同时, 为提高分类器的泛化能力, 研究者们还将集成学习引入到文本分类中. Weiss等[2]用决策树的集成进行文本分类, 并且成功地用于 email的过滤. Schapire等[3]将决策树桩 (decision stump) 的集成用于文本分类系统 Boostexter中, 也取得了较好的效果. 这些算法在一定程度上提高了文本分类的性能. 然而, 这些算法主要是从分类器设计的角度出发, 通过某种优化策略来改进分类器的性能, 而采用的文本表示为传统的向量空间模型 (Vector Space Model, VSM), 未能有效地利用文本中的结构信息.

VSM 采用一个向量表示一个文档[4], 仅考虑单词在文本中出现的频率, 未考虑不同位置的单词表达文本内容的差异性, 因而在某些情况下使得基于 VSM 的文本相似性度量的分类性能下降, 如 "here you are"和 "you are here"两个句子, 由于单词的摆放位置不同, 使得句子的含义大相径庭. 此外, 一个单词出现在标题中和出现在文档正文中甚至出现在段首、段尾, 对整篇文档来说其重要性是不同的, 即单词在文档

收稿日期: 2009-05-05

基金项目: 国家自然科学基金 (60873176)、江苏省自然科学基金 (BK2008430)资助项目.

通讯联系人: 杨 明, 教授, 博士生导师, 研究方向: 数据挖掘, 机器学习, 模式识别, 粗集理论. E-mail: m. yang@ njnu. edu. cn

中的出现位置,将对文本的分类效果有很大的影响.

近年来,多模态数据融合已成为数据挖掘的重要研究领域.文献[5]在考虑视频中的多种信息流如视觉流、听觉流以及文本流的基础上,提出了一种基于多模态组合的视频事件索引方法.文献[6]提出了时间间隔多媒体事件(time intervalsmultimedia events,TIME)的概念,TIME通过考虑视频流中的上下文和同步关系,将多种异构信息源结合起来共同表达某一语义事件.文献[7]从视频语义特征的角度出发,将与视频语义相关的声音、字幕、音乐、剧情脚本、新闻文稿等信息特征进行整合,通过人像、字幕、语音、视频镜头识别和剧情脚本分析等组合手段,实现视频语义特征的多模式提取和检索.文献[8]将文本、视觉和听觉等多模态数据构成三阶张量,根据视频的时序共生特性设计了"张量镜头"的降维方法,并根据"张量镜头"提出了直推式支持张量机算法,来实现视频镜头的语义概念检测.

为有效利用文本中单词的结构信息,文献[9]提出了"分布特征"的概念.本文在其基础上引入在视频检测中经常使用的多模态的概念,将单词在文本中出现的平均位置和标准差这两种模态嵌入到文本表示中,提出了一种多模态向量空间模型(Multi-Modality Vector Space Model,MMVSM);在此基础上,给出了一种新的文本相似性度量 MMSM.进一步,为有效降低抽取的特征向量维数,本文引入了基于随机子空间的分类集成方法,提出了一种多模态模型的随机子空间多分类器集成算法(Multi-Modality-based Random Feature subspace classifier Ensemble,MMRFSEn),该算法为确保集成分类器独立性,运用特征子集来构建各个体分类器,以此来提高分类器的泛化能力.实验结果表明,本文提出的 MMRFSEn算法是有效可行的.

1　基于多模态模型的相似性度量 MMSM

传统的文本表示方法采用 VSM 模型,即对某个文本 D_i 用向量 $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \cdots, d_{in})$ 来表示,其中 $d_{ik}(k = 1, 2, \cdots, n)$ 为相应单词在 D_i 中的权重,即 tf-idf 值,其计算公式如下:

$$d(t_k, D_k) = \frac{tf(t_k, D_k) \times \log\left(\frac{M}{df(t_i)} + 0.01\right)}{\sqrt{\sum_{j=1}^{n}\left[tf(t_j, D_k) \times \log\left(\frac{M}{df(t_j)} + 0.01\right)\right]^2}}. \tag{1}$$

式中,$tf(t_k, D_k)$ 表示单词 t_i 在文本 D_k 中出现的频率,$df(t_i)$ 表示特征项的文本频,M 表示总的文本数,n 表示特征总数.

对给定的两个文档 D_i 和 D_j,相应的向量分别为 $\boldsymbol{d}_i$ 和 $\boldsymbol{d}_j$,D_i 和 D_j 的相似度就可以借助于向量之间的某种距离来表示,其中经典的度量方法采用向量之间夹角的余弦函数来计算,其公式如下:

$$SM(\boldsymbol{d}_i, \boldsymbol{d}_j) = \cos(\boldsymbol{d}_i, \boldsymbol{d}_j) = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{|\boldsymbol{d}_i| \cdot |\boldsymbol{d}_j|}. \tag{2}$$

可以看出,根据式(2),"here you are"和"you are here"这两个句子是完全相似的,显然这不符合实际应用的需要.造成度量失准的主要原因是:该度量策略未能有效地利用各单词的结构信息.同理,对于两个文档,若其各单词出现的次数相同,则依据式(2),它们是完全相似的,这显然不符合人们的直觉.直觉上,如果某单词在两个文档中出现的次数相同,并且出现的平均位置也很相近,那么这个单词对于两个文档的相似性具有较高的贡献,否则若仅次数相同,但平均位置相差很大,那么该单词的贡献相对较小,应受到一定程度的抑制.因此,本文为每个单词都设置一个抑制因子 w 来控制结构信息对文本相似性的影响.同时,为充分考虑文本的结构信息,本文采用权值向量,以及文本中各单词出现的平均位置及标准差 3个模态来表示一个文本.为方便计,将文本的这种表示简称为多模态向量空间模型(Multi-Modality Vector Space Model,MMVSM).

对给定的文本 D_i,其 3组向量分别为 $\boldsymbol{d}_i = \{v_{i1}, \cdots, v_{in}\}$,$\mu_i = \{\mu_{i1}, \cdots, \mu_{in}\}$,$\sigma_i = \{\sigma_{i1}, \cdots, \sigma_{in}\}$,其中 $\boldsymbol{d}_i$ 为文本中各单词在 D_i 中的权值组成的向量,μ_i 和 σ_i 分别是相应单词的平均位置和标准差组成的向量,n 是向量的维数.有关单词的平均位置和标准差的计算策略[10]见式(3)和(4).

设文本 D 中有 n 个句子,单词 t 在每个句子中出现的次数为 c_i,则单词出现的平均位置 $\mu(t, d)$ 和标准差 $\sigma(t, d)$ 分别为:

$$\mu(t, d) = \frac{\sum_{i=1}^{n} c_i \times i}{\mathrm{count}(t, d)}, \tag{3}$$

$$\sigma(t, d) = \sqrt{\sum_{i=1}^{n} \frac{c_i \times (i - \mu(t, d))^2}{\mathrm{count}(t, d)}}. \tag{4}$$

式中, $\mathrm{count}(t, d)$表示单词 t在文本 d中出现的总次数. 为有效利用文本中单词的结构信息, 本文提出一种基于MMVSM的文本相似性度量策略, 具体方法为: 设文档 D_i和D_j为给定的两个文本, 其中D_i的 3组向量分别为 $\boldsymbol{d}_i = \{v_{i1}, \cdots, v_{in}\}$, $\mu_i = \{\mu_{i1}, \cdots, \mu_{in}\}$, $\sigma_i = \{\sigma_{i1}, \cdots, \sigma_{in}\}$, D_j的 3组向量分别为 $\boldsymbol{d}_j = \{v_{j1}, \cdots, v_{jn}\}$, $\mu_j = \{\mu_{j1}, \cdots, \mu_{jn}\}$, $\sigma_j = \{\sigma_{j1}, \cdots, \sigma_{jn}\}$, n是向量的维数. 为有效地将文本中单词的位置和标准差这些结构信息嵌入到相似性度量中, 本文提出一种新的基于 MMVSM 的文本相似性度量策略 MMSM, 其函数表示如下:

$$\mathrm{MMSM}(\boldsymbol{d}_i, \boldsymbol{d}_j) = \frac{\sum_{k=1}^{n} w_{ijk} \times v_{ik} \times v_{jk}}{\sqrt{\sum_{k=1}^{n} v_{ik}^2} \times \sqrt{\sum_{k=1}^{n} v_{jk}^2}}, \tag{5}$$

其中,

$$w_{ijk} = \begin{cases} w, & |\mu_{ik} - \mu_{jk}| > \overline{\mu_k}, \ |\sigma_{ik} - \sigma_{jk}| < \overline{\sigma_k}, \quad 0 < w < 1 \\ 1, & \text{否则} \end{cases} \tag{6}$$

$$\overline{\mu_k} = \frac{\sum_{i=1}^{M} \mu_{ik}}{M}, \tag{7}$$

$$\overline{\sigma_k} = \frac{\sum_{i=1}^{M} \sigma_{ik}}{M}, \tag{8}$$

式中, M 为训练集的样本个数. 在新的相似性度量下, 两个文本之间的距离表示为:

$$\mathrm{Dis}(\boldsymbol{d}_i, \boldsymbol{d}_j) = 1 - \mathrm{MMSM}(\boldsymbol{d}_i, \boldsymbol{d}_j). \tag{9}$$

由式 (6)可以看出, 若单词T在文档D_i和D_j中出现的平均位置相差很大, 而标准差相差在一个很小的范围内, 则表明 T 的结构信息对于这两篇文档的相似性贡献相对较小, 它的作用应该受到一定程度的抑制, 即抑制因子 w 小于 1, 否则, 抑制因子 w为 1, 此时表示两个文档在结构信息上很相似, 只考虑单词的权值, 多模态模型退化为标准的 VSM 模型. 可见, 基于多模态向量空间模型充分考虑了每个单词的结构信息对于文档相似性的影响, 是基于VSM的相似性度量的推广和扩展, 也是基于VSM相似性度量的有效改进.

2 基于多模态模型的随机子空间分类集成算法 *MMRFSEn*

为有效利用单词的结构信息以及集成学习的泛化能力, 本文提出了基于多模态模型的随机子空间分类集成算法 (MMRFSEn), 如图 1所示. 本文中子空间的选择是根据均匀分布 U 随机抽取 m 个不同的子集 $A = \{d_1, d_2, \cdots, d_m\}$, 每个子集的大小 (即子空间的维数)为 r, 并且每个子集包含 3个模态, 即权值、平均位置和标准差. 每个子空间都定义一个映射 $P_A: F^n \rightarrow F^m$, 在此基础上得到每个训练子集 $D_i = \{(P_A(x_j), y_j) \mid 1 \leqslant j \leqslant N\}$. 再由分类算法 L训练该数据子集并得到待检样本的决策 h_i, 重复 m 次, 最后利用择多投票法得到待检样本的最终决策. 在计算文本相似度时, 采用基于多模态模型的相似性度量 MMSM.

事实上, 每个子空间的维数 r和子空间的个数 m 可自动确定. 为简便, 本文采用事先设定的方法, 即给定相对固定的值. 本文实验中, 分别采用了在文本分类中常用的 SVM、KNN、RBFNetwork和 NaiveBayes 4种分类器作为基分类器, 进行同态集成.

依据上述分析, 基于多模态模型的随机子空间分类集成算法 (MMRFSEn) 的具体步骤描述如下:

输入: (1) 训练集 $D = \{(x_j, y_j) \mid 1 \leqslant j \leqslant N\}$, $x_j \in X \subset \mathbf{R}^n$, $y_j \in \mathbf{C} = \{1, \cdots, k\}$;

(2) 学习算法 (训练基分类器)L;

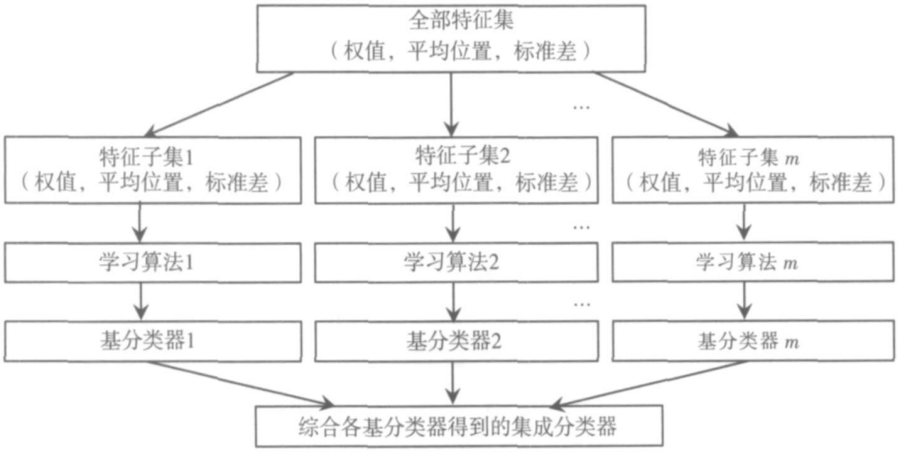

图 1 基于多模态模型的随机子空间分类集成

Fig.1 Multi-modality random subspace classifier ensemble

(3) 子空间维数 $r < n$;

(4) 子空间的个数 m;

输出: 集成分类器 H.

步骤:

(1) for($i = 1$; $i <= m$; $i++$)

随机产生 r维向量 ***select***$_i$, 保存该子空间在原始空间中的位置;

(2) 对每一个特征子集为 ***select***$_i$, 在训练集 D 上进行投影得到该子空间上的训练数据集 D_i (P_i, U_i, V_i), 其中 P_i、U_i、V_i 分别表示频率、平均位置和标准差在第 i个子集上的子数据集;

(3) 由分类算法 L和各个子空间上的多模态训练数据 D_1, …, D_m 来训练分类器 classifier$_1$, classifier$_2$, …, classifier$_m$;

(4) 对于待检样本 X_i, 抽取该样本的多模态模型, 然后根据特征子集 ***select***$_1$, …, ***select***$_m$ 得到该样本对应的多模态子样本 X_{i1}, …, X_{im};

(5) $H(X_i) = \underset{y \in C}{\arg\max} \sum_{j\ \mathrm{Classifier}_j(X_{ij})=y} 1$

基于多模态随机子空间的分类集成算法 (MMRFSEn), 该算法可有效地利用单词出现的频率及单词在文本中的结构信息, 因而可提高文本分类的性能. 此外, 在降低特征空间维数的同时, 基于不同特征子空间构建基分类器策略可有效降低分类器之间的相关性, 且可避免直接在原始空间上进行特征选择带来的信息丢失问题, 可改进集成分类器的性能.

3 实验结果及分析

为验证 MMRFSEn算法的有效性, 本文采用 20Newsgroup、WebKB、Reuters 3个数据集进行实验验证.

3.1 数据集

20Newsgroups数据集包含从新闻组收集到的 19 997个文章, 这里使用了其中的 5类共 5 000个文档; WebKB数据集包含着从 4个学校获得的 8 282个 Web网页, 本文选择其中的 4 000个网页; Reuters数据集包含 21 578个取自路透社新闻专线的文章, 本文采用的是“ModApte”版本的 Reuters-21578文集, 即从原始的文集中滤去没有标记的文档, 并且只选择那些至少在训练集和测试集中各出现一个文档的那些类别来组成文集, 并在样本数最多的 5个类别上进行了实验.

3.2 分类性能的评估

在本文的实验中, 使用的性能评估指标为常用的 F1-value, 为了更好地介绍查全率和查准率以及 F1-value的含义, 建立以下混合矩阵, 如表 1表示. 混合矩阵是针对某一个类而言的, 它统计了所有测试文本与这个待定类之间的分类情况.

表 1 混合矩阵

Table 1 Confusion matrix

	实际属于该类的文本数	实际不属于该类的文本数
分类器认为属于该类	***TP***	***FP***
分类器认为不属于该类	***FN***	***TN***

则查全率 (Recall)和查准率 (Precision)以及 F1-value定义如下:

$$\text{Recall}=\frac{TP}{TP+FN}, \tag{10}$$

$$\text{Precision}=\frac{TP}{TP+FP}, \tag{11}$$

$$\text{F1-value}=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}. \tag{12}$$

对于多类别的问题,一般采用平均的方法: 微平均 (micro-average)和宏平均 (macro-average). 微平均统一计算全部类别的召回率、准确率和 F1值,即从整体上平均. 宏平均计算每一类的召回率、准确率和 F1值后取算术平均值. 宏平均值更多地受到稀有类别 (包含文档较少或出现概率较小的类别)的影响.

3.3 实验结果分析

本文使用 Weka软件中的 SVM、RBFNetwork、NaiveBayes、KNN作为基本分类器,并作适当地修改,在上述 3个数据集上做了实验. 实验中,原始空间的维数为 900,子空间的维数取 500,集成的大小为 15,KNN算法中 k取 20,w取 0.7. 表 2给出了 MMRFSEn算法和采用传统的 VSM模型的子空间分类器集成算法 RFSEn以及两种模型下的单一分类器的实验结果. 其中,每个数据集上最好的结果用粗体表示. 此外,图 2~5给出 w的取值变化对在不同数据集和不同基分类器下的 mic-$f1$值的影响.

表 2 分类性能比较

Table 2 Classification performance comparison

		SVM		NaiveBayes		RBFNetwork		KNN	
		mac-$f1$	mic-$f1$	mac-$f1$	mac-$f1$	mic-$f1$	mic-$f1$	mac-$f1$	mic-$f1$
Newsgroup	RFSEn	0.9240	0.9249	0.8639	0.8576	0.8659	0.8677	0.8867	0.8885
	MMRFSEn	**0.9301**	**0.9318**	**0.8659**	**0.8686**	**0.8761**	**0.8810**	**0.8956**	**0.9013**
	MMSingle	0.9204	0.9215	0.8642	0.8680	0.8662	0.8731	0.8933	0.8951
	Single	0.9144	0.9154	0.8629	0.8520	0.8526	0.8565	0.8926	0.8940
Reuters	RFSEn	0.8505	0.8618	0.7634	0.7764	0.7237	0.7422	0.8176	0.8345
	MMRFSEn	**0.8515**	**0.8640**	0.7789	0.7820	**0.7442**	**0.7436**	0.8220	0.8354
	MMSingle	0.8488	0.8629	**0.7801**	**0.7850**	0.7323	0.7361	**0.8236**	**0.8369**
	Single	0.8478	0.8602	0.7684	0.7810	0.7080	0.7236	0.8198	0.8338
WebKB	RFSEn	0.8834	0.8967	0.7849	0.7929	0.7556	0.7710	0.7123	0.7498
	MMRFSEn	**0.8896**	**0.9029**	0.7910	0.8115	**0.7601**	**0.7812**	0.7168	0.7506
	MMSingle	0.8836	0.8928	**0.8024**	**0.8135**	0.7421	0.7564	**0.7286**	**0.7589**
	Single	0.8755	0.8884	0.7933	0.8012	0.7337	0.7530	0.7098	0.7467

图 2和图 3分别给出了以 KNN为基分类器时,在不同的抑制程度 w的影响下,MMSingle算法和 MMRFSEn算法在 WebKB和 Reuters上的分类效果.

图 4和图 5分别给出了以 NaiveBayes为基分类器时,在不同的抑制因子 w的影响下,MMSingle算法和 MMRFSEn算法在 WebKB和 Reuters上的分类效果.

从表 2可以看出,多数情况下,MMRFSEn算法的效果要优于传统的 VSM模型下的集成效果. 进一步,由图 2和图 3可以看出,采用 KNN为基分类器时,在 WebKB和 Reuters数据集上,当 $w=0.6$和 $w=0.5$时,集成的效果最好; 而由图 4和图 5可知,以 NaiveBayes为基分类器时,当 $w=0.8$和 $w=0.9$时,集成的效果最好. 可见,w的取值对集成分类器的性能有很大的影响,这也说明表 2中在 WebKB和 Reuters数据集上,MMRFSEn算法以 KNN算法或 NaiveBayes算法作为基分类器时集成性能不佳的主要原因,可能与 $w=0.7$有关. 这也表明表 2中,由 MMRFSEn算法得到的集成分类器不是优化的结果,如何自适应选择 w的值将是未来研究的主要工作之一.

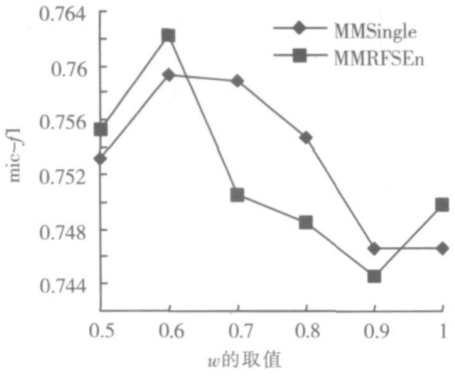

图 2　WebKB 上 w 值的变化对 KNN 的影响

Fig.2　The influence on KNN with w on WebKB

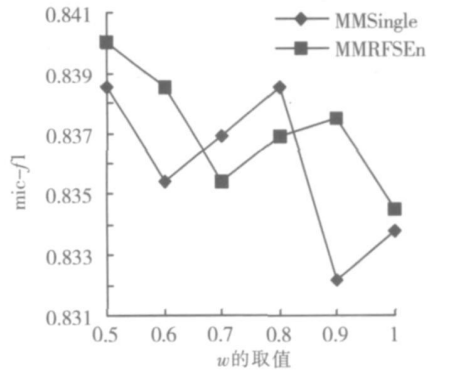

图 3　Reuters 上 w 值的变化对 KNN 的影响

Fig.3　The influence on KNN with w on Reuters

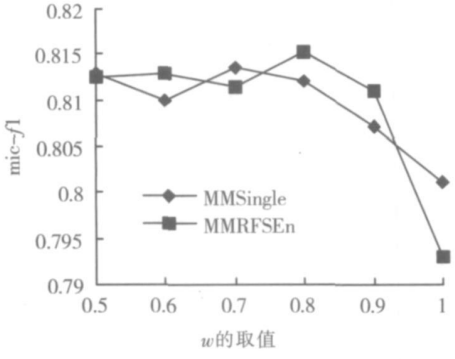

图 4　WebKB 上 w 值的变化对 NaiveBayes 的影响

Fig.4　The influence on NaiveBayes with w on WebKB

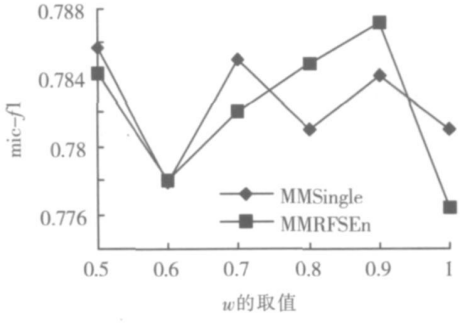

图 5　Reuters 上 w 值的变化对 NaiveBayes 的影响

Fig.5　The influence on NaiveBayes with w on Reuters

4　结语

本文引入一种多模态向量模型 MMVSM,并提出了一种基于多模态模型的相似性度量方法,该方法可合理使用单词的结构信息,因而有效改进和拓展传统的文本相似性度量. 基于 MMVSM,结合随机子空间的分类集成,提出了一种基于多模态模型的随机子空间分类集成算法 MMRFSEn,并在标准数据集上,与 VSM 模型下的 RFSEn算法及单一分类器进行了实验比较. 实验结果表明,MMRFSEn算法可有效提高文本分类的性能,改进文本分类器的推广性,为文本的表示和分类提供了一种新的途径.

下一步的研究工作将着重于设计特征子空间的优化策略且可自适应选择 w 值的分类器集成算法.

[参考文献] (References)

[1] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Survey, 2002, 34(1): 1-47.

[2] Weiss S M, Apte C, Damerau F J. Maximizing text-mining performance[J]. IEEE Intelligent Systems, 1999, 14(4): 63-69.

[3] Schapire R E, Singer Y. Boostexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(223): 135-168.

[4] Lu Yuchang, Lu Mingyu, Li Fan. Annlysis and construction of word weighing function in VSM[J]. Journal of Computer Research & Development, 2002, 39(10): 1 205-1 210.

[5] Babaguchi N, Kawai Y, Kitahashi T. Event based indexing of broadcast sports video by intermodal collaboration[J]. IEEE Trans on Multimedia, 2002, 4(1): 68-75.

[6] Snoek C G M, Worring M. Multimedia event-based video indexing using time intervals[J]. IEEE Trans on Multimedia, 2005, 7(4): 638-647.

[7] Hu N, Wang Y W, Lü N. Study on multimodel retrieval method of content-based video[J]. Journal of Jilin University: Information Science Edition, 2006, 24(3): 265-270. (in Chinese with English abstract).

(下转第 72页)

点移动速度不是很快时,频繁的计算路由代价需要消耗大量不必要的能量;对于边界模式,平面化网络过程以及产生的过多的跳跃往往消耗较多的能量;在考虑路由选择时没有考虑节点的剩余能量,往往会导致某些节点能量的快速消耗代价,造成较多空洞情况,形成恶性循环,减短了网络的寿命.

本文提出一种简易的、更有效的策略,采取 GEAR中把节点剩余能量作为路由选择考虑因素的方法,节点存储带有标记的单跳路由状态表,对 GPSR的路由转发机制进行了改进,减少了由 GPSR发现的路由造成的网络寿命减短;同时当转发节点知道自己为边界转发节点时通过发送警告信息通知邻居节点,避免了信息进入同一个空洞中,对 GPSR协议做了完善和改进.由于单跳路由有效时间 T、距离对路由选择的贡献或者影响大小 α以及剩余能量对路由选择的贡献和影响大小 β的值需要根据实际情况和主观经验确定,今后的工作是通过仿真软件对这些值给出比较好的参考选择范围.

[参考文献](References)

[1] 郑少仁,王海涛,赵志峰,等. Ad Hoc网络技术[M]. 北京:人民邮电出版社,2005.
Zheng Shaoren, Wang Haitao, Zhao Zhifeng, et al. Ad hoc Network[M]. Beijing: Posts&Telecom Press, 2005. (in Chinese)

[2] 牛新征,周明天,佘堃. 一种新的移动自组网的数据传输策略[J]. 计算机应用研究,2009,26(2):660-664.
Niu Xinzheng, Zhou Mingtian, She Kun. New data transmission scheme in mobile Ad hoc networks[J]. Application Research of Computers, 2009, 26(2): 660-664. (in Chinese)

[3] 郑锴,童利标,陆文骏. 基于地理位置的无线传感器网络路由协议[J]. 中兴通讯技术,2008,14(6):37-41.
Zheng Kai, Tong Libiao, Lu Wenjun. Routing algorithms based on location information for wireless sensor network[J]. ZTE Communications, 2008, 14(6): 37-41. (in Chinese)

[4] Karp B, Kung H T. Greedy perimeter stateless routing for wireless networks[C]//Pro of the Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking(MobiCOM 2000). Boston, 2000: 243-254.

[5] Yu Yan, Govindan R, Estrin D. Geographical and energy-aware routing: A recursive data dissemination protocol for wireless sensor networks[R]. UCLA Computer Science Department, 2001: 1-23.

[6] 张耀,贾振红. 求解路由空洞问题的 GEAR改进算法[J]. 计算机工程,2008,34(12):94-95.
Zhang Yao, Jia Zhenhong. Improved GEAR algorithm for solving routing hole problem[J]. Computer Engineering, 2008, 34(12): 94-95. (in Chinese)

[责任编辑:严海琳]

(上接第 62页)

[8] 吴飞,刘亚楠,庄越挺. 基于张量表示的直推式多模态视频语义概念检测[J]. 软件学报,2008,19(11):2 583-2 868.
Wu Fei, Liu Yanan, Zhuang Yueting. Transductive multi-modality video concept detection with Tensor representation[J]. Journal of Software, 2008, 19(11): 2 583-2 868. (in Chinese)

[9] Xue Xiaobing, Zhou Zhihua. Distributional features for text categorization[C]//Proceedings of the 17th European Conference on Machine Learning(ECML'06). Berlin, Germany: LNAI 4212, 2006: 497-508.

[10] 孙春红,杨明. 一种嵌入分布信息的 Web文档相似性度量[J]. 南京师范大学学报:工程技术版,2008,8(3):67-68.
Sun Chunhong, Yang Ming. A novel similarity measurement for web pages by incorporating distribution information[J]. Journal of Nanjing Normal University: Engineering and Technology Edition, 2008, 8(3): 67-68. (in Chinese)

[责任编辑:严海琳]