基于图正则化 MNMF 的中文垃圾邮件过滤*

刘遵雄[1]，黄志强[1†]，郑淑娟[2]，石　菲[1]

（1. 华东交通大学 信息工程学院，南昌 330013；2. 江西财经大学 科研处，南昌 330013）

摘　要：利用向量空间模型表示的文本邮件数据具有高维性，不利于邮件过滤模型的建立，需要对数据进行降维处理。最大间隔 Semi-NMF（max-margin semi-nonnegative matrix factorization，MNMF）能够同时实现维数约减和邮件分类，而图正则化 NMF 能保持数据空间的几何结构。基于以上两种 NMF 改进模型，提出了图正则化 MNMF（graph regularized MNMF，GMNMF）算法，并设计了一个迭代的求解算法。将 GMNMF 算法及其他相关算法用于中文垃圾邮件过滤实验，结果表明 GMNMF 算法构建的过滤模型要优于其他较好的算法构建的过滤模型。

关键词：向量空间模型；维数约减；最大间隔 Semi-NMF；图正则化 MNMF；中文垃圾邮件过滤

中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1001-3695(2013)09-2672-05

doi:10.3969/j.issn.1001-3695.2013.09.028

Chinese spam filtering based on graph regularized MNMF algorithm

LIU Zun-xiong[1]，HUANG Zhi-qiang[1†]，ZHENG Shu-juan[2]，SHI Fei[1]

（1. *School of Information Engineering*，*East China Jiaotong University*，*Nanchang* 330013，*China*；2. *Division of Scientific Research*，*Jiangxi University of Finance & Economics*，*Nanchang* 330013，*China*）

Abstract：Text e-mail data represented by vector space model（VSM）are high dimensionality. This situation is not conducive to construct e-mail filtering model. Therefore，dimensionality reduction need be performed. MNMF could simultaneously achieve dimensionality reduction and e-mail classification，and graph regularized NMF could keep the geometrical structure of the data space. Based on the above two improved NMF models，this paper put forward GMNMF algorithm，and designed an iterative solution algorithm. Using GMNMF algorithm and other related algorithm do Chinese spam filtering experiments. The experimental results show that the model of the proposed algorithm is superior to models of other good algorithms.

Key words：vector space model（VSM）；dimensionality reduction；MNMF；GMNMF；Chinese spam filtering

0　引言

电子邮件作为网络提供的服务之一，因其具有方便、快捷、经济等优点而被广泛地用于交流和通信。然而，垃圾邮件的大量出现给企业和个人带来了许多烦恼。它不仅占用大量的网络资源，浪费用户的时间和精力，影响网络服务提供商的形象，而且很有可能包含病毒以及一些非法有害内容[1]。因此，研究快速、准确、智能的垃圾邮件过滤方法具有十分重要的实际应用价值。

为了应对大量出现的垃圾邮件，各种垃圾邮件过滤方法被相继提出，其中基于内容的垃圾邮件过滤方法是当前比较主流的方法。基于内容的过滤方法把垃圾邮件过滤归结为文本分类问题。许多相关研究表明[1~3]，在进行邮件分类之前使用一些特征抽取方法能改善垃圾邮件过滤的性能。常见的特征抽取方法有主成分分析（principal component analysis，PCA）、线性判别分析（linear discriminant analysis，LDA）、独立成分分析（independent component analysis，ICA）、非负矩阵分解（non-negative matrix factorization，NMF）等[4,5]，且比较典型的过滤算法包括贝叶斯（Naïve Bayesian，NB）、支持向量机（support vector machine，SVM）、逻辑回归（logistic regression，LR）等[6~8]。

NMF 是经典的机器学习算法之一，能够很好地实现原数据的低维近似，并且已被广泛地应用于计算机视觉[9]、模式识别[10]等许多领域中。Ding 等人[11]放松了 NMF 中的数据矩阵和基矩阵的非负约束，提出了 Semi-NMF 算法，并且设计了一个迭代更新算法用于求解它。为了同时进行矩阵分解和分类，Kumar 等人[12]将最大间隔框架引入到 Semi-NMF 中，提出了 MNMF 算法，还给出了一个迭代更新求解算法，并且通过对比实验证明了其在分类精度等方面优于 Semi-NMF + SVM 和 DNMF[13]。针对数据空间的几何结构，Cai 等人[14]针对 NMF 构造了一个最近邻图，形成了图正则化 NMF（graph regularized non-negative matrix factorization，GNMF）算法并且给出了简单的更新规则。据此，本文提出了 GMNMF 算法，同时相应地设计了一个求解算法，将 GMNMF 算法用于构建中文垃圾邮件过滤模型。结合两个公开的中文邮件数据集（CDSCE[15] 和 trec06c[16]）进行仿真实验，结果表明经 GMNMF 特征抽取后的邮件数据更容易分类，并且 GMNMF 算法构建的模型要优于 MNMF、NMF + SVM、PCA + SVM、SVM 等构建的模型。

1　非负矩阵分解及其各种变形

从文本分类的角度来看，垃圾邮件过滤问题是一个二值分

收稿日期：2013-01-10；**修回日期**：2013-03-02　　**基金项目**：国家自然科学基金资助项目(61065003)；国家教育部人文社会科学研究规划基金资助项目(10YJC630379)

作者简介：刘遵雄(1967-)，男，江西瑞昌人，教授，硕导，博士，主要研究方向为机器学习、数据挖掘；黄志强(1989-)，男，(通信作者)，江西临川人，硕士研究生，主要研究方向为机器学习及优化算法(huangylqf@gmail.com)；郑淑娟(1971-)，女，江西宜黄人，馆员，硕士，主要研究方向为数据统计分析与推断；石菲(1988-)，女，安徽安庆人，硕士研究生，主要研究方向为模式识别.

类问题:一类是垃圾邮件(spam e-mail),标记为 -1;另一类是正常邮件(legitimate e-mail 或 ham e-mail),标记为 1。假设给定中文邮件数据集的特征矩阵 $X=[x_1,\cdots,x_i,\cdots,x_n]\in\mathbb{R}^{m\times n}$ 以及相应的类别矩阵 $Y=[y_1,\cdots,y_i,\cdots,y_n]\in\mathbb{R}^{1\times n}$。其中 m 表示每一封邮件的特征维数,n 表示邮件的总数,$x_i\in\mathbb{R}^m$ 且 $y_i\in\{-1,1\}$,并使用向量空间模型(VSM)[17]来描述文本邮件数据,所以特征矩阵 X 是一个非负矩阵。下面就将对 NMF 算法及其各种变形作一个简单的介绍。

1.1　非负矩阵分解

NMF 算法就是对于上述给定的非负矩阵 X,要找到两个非负矩阵 U 和 V 使得 $X\approx UV$,其中基矩阵 $U\in\mathbb{R}^{m\times k}$ 且系数矩阵 $V\in\mathbb{R}^{k\times n}$。为了达到维数约减的目的,通常选取的 k 要比 m 和 n 小得多。通过观察上面的近似等式可以发现,$x_i\approx Uv_i$,其中 $x_i\in\mathbb{R}^m$ 和 $v_i\in\mathbb{R}^k$ 分别为 X 和 V 对应的列向量。因此,特征矩阵 X 的每一列都可用基矩阵 U 的所有列的线性组合来近似表示,并且对应的坐标为系数矩阵 V 的相应列。

为了找到基矩阵 U 和系数矩阵 V,本文使用如下代价函数(即最小化重构误差):

$$\arg\min_{U,V}\|X-UV\|_F^2\qquad \text{s.t. } U\geqslant 0,V\geqslant 0 \tag{1}$$

其中:$U\geqslant 0$ 和 $V\geqslant 0$ 分别表示矩阵 U 和 V 的每一个元素都不小于 0,$\|\cdot\|_F$ 表示 Frobenius 范数[18]。观察式(1)的优化问题可以看出:其仅是 U 或者 V 的凸函数,而不是 (U,V) 的凸函数,所以要想找到它的全局最优解就较为困难。为了找到上述问题的局部最优解,Lee 等人[19]提出了下面的乘性迭代规则:

$$V\leftarrow V\odot\frac{U^{\mathrm{T}}X}{U^{\mathrm{T}}UV},U\leftarrow U\odot\frac{XV^{\mathrm{T}}}{UVV^{\mathrm{T}}} \tag{2}$$

其中:$\odot$ 表示矩阵按元素乘。

1.2　NMF 的各种变形

随着对 NMF 算法的不断深入研究,NMF 算法的不足之处也逐渐地被发现,改进的 NMF 算法也就被相继提出,但大多是基于放松非负约束或者向代价函数引入其他惩罚项的改进。Ding 等人[11]放松了 NMF 算法对 X 和 U 的非负约束,得到了 Semi-NMF 算法:

$$\arg\min_{U,V}\|X-UV\|_F^2\qquad \text{s.t. } V\geqslant 0$$

为了使矩阵分解和分类能够同时进行,Kumar 等人[12]将最大间隔框架引入到 Semi-NMF 中,提出了 MNMF 算法:

$$\arg\min_{U,V,w,b,\varepsilon}\|X-UV\|_F^2+\lambda\left(\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{n}\varepsilon_i\right)$$
$$\text{s.t.}\quad V\geqslant 0$$
$$y_i(w^{\mathrm{T}}U^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i\quad \varepsilon_i\geqslant 0,1\leqslant i\leqslant n$$

其中:$\varepsilon\in\mathbb{R}^n$,$\lambda>0$ 是正则化参数,超平面参数 $w\in\mathbb{R}^k$ 且 $b\in\mathbb{R}$。

为了维持数据空间的几何结构,Cai 等人[14]针对 NMF 算法构造了一个最近邻图,如此便形成了 GNMF 算法:

$$\arg\min_{U,V}\|X-UV\|_F^2+\lambda Tr(VLV^{\mathrm{T}})\qquad \text{s.t.}\quad U\geqslant 0,V\geqslant 0$$

其中:λ 与 MNMF 算法中的一样;$Tr(\cdot)$ 表示矩阵的迹;$L\in\mathbb{R}^{n\times n}$ 是图拉普拉斯(graph Laplacian)矩阵。GNMF 算法比较适合用于聚类分析,但不能利用已知的类别信息,因此不太适合于分类问题。

针对 NMF 算法不能利用已知的类别信息,Liu 等人[20]提出了约束非负矩阵分解(constrained nonnegative matrix factorization,CNMF)算法:

$$\arg\min_{U,Z}\|X-UZ^{\mathrm{T}}A^{\mathrm{T}}\|_F^2\qquad \text{s.t.}\quad U\geqslant 0,Z\geqslant 0$$

其中:Z 是辅助矩阵,A 是类别约束矩阵。上述各种改进的 NMF 的具体求解算法可参考对应的相关文献。

2　图正则化 MNMF

为了保证维数约减和邮件分类能够同时进行以及能够保持数据空间的几何结构,本文将一个最近邻图引入到 MNMF 算法中,从而提出了图正则化 MNMF(GMNMF)算法,并且就该算法的求解设计了一个求解算法。为了避免计算复杂,本文仅根据邮件是否同类来判断它们对应的数据点是否为近邻点(若 y_i 和 y_j 同类,则 x_i 和 x_j 是近邻点;反之亦然),同时选择 0-1权重[14]来得到权重矩阵 W。依据一般的正则化框架,有 GMNMF 算法的代价函数为

$$\arg\min_{U,V,w,b,\varepsilon}\|X-UV\|_F^2+\lambda_1\left(\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{n}\varepsilon_i\right)+\lambda_2 Tr(VLV^{\mathrm{T}})$$
$$\text{s.t.}\quad V\geqslant 0$$
$$y_i(w^{\mathrm{T}}U^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i\quad \varepsilon_i\geqslant 0,1\leqslant i\leqslant n \tag{3}$$

其中:λ_1、λ_2 和 C 都是大于零的常数;$L=D-W$,D 是一个对角矩阵且 $D_{ii}=\sum_{j=1}^{n}W_{ij}(i=1,\cdots,n)$;权重矩阵 W 为对称矩阵。观察式(3)的优化问题可知,直接求解会比较困难,因此本文按照文献[12]的思路(即在求解某一或几项时,仅更新要求解的而保持其他变量值固定不变)设计了下面的迭代更新求解算法:

a)为了更新基矩阵 U,需要保持 V、w、b 固定不变。此种情况下,式(3)转换成如下的优化问题:

$$\arg\min_{U,\varepsilon}\|X-UV\|_F^2+\lambda_1 C\sum_{i=1}^{n}\varepsilon_i$$
$$\text{s.t.}\quad y_i(w^{\mathrm{T}}U^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i\quad \varepsilon_i\geqslant 0,1\leqslant i\leqslant n \tag{4}$$

针对式(4)的有约束优化问题,引入拉格朗日乘子 $\alpha_i\geqslant 0$ 和 $\beta_i\geqslant 0(i=1,\cdots,n)$,得到对应的拉格朗日函数:

$$l(U,\varepsilon,\alpha,\beta)=\|X-UV\|_F^2+\lambda_1 C\sum_{i=1}^{n}\varepsilon_i-$$
$$\sum_{i=1}^{n}\alpha_i[y_i(w^{\mathrm{T}}U^{\mathrm{T}}x_i+b)-1+\varepsilon_i]-\sum_{i=1}^{n}\beta_i\varepsilon_i \tag{5}$$

对式(5)中的主变量 (U,ε) 分别求偏导数且令它们分别等于零,有

$$\frac{\partial l}{\partial U}=0\Rightarrow U=(2XV^{\mathrm{T}}+\sum_{i=1}^{n}\alpha_i y_i x_i w^{\mathrm{T}})(2VV^{\mathrm{T}})^{-1}$$
$$\frac{\partial l}{\partial \varepsilon_i}=0\Rightarrow\lambda_1 C-\alpha_i-\beta_i=0\quad 0\leqslant\alpha_i\leqslant\lambda_1 C,i=1,\cdots,n \tag{6}$$

将式(6)代入到式(5)中,同时根据有关的线性代数知识,可以得到原优化问题的对偶问题:

$$\arg\max_{\alpha}\alpha^{\mathrm{T}}(T_1-T_2)\alpha+(t_3-t_4-t_5-t_6+t_7)\alpha$$
$$\text{s.t.}\quad 0\leqslant\alpha_i\leqslant\lambda_1 C,i=1,\cdots,n \tag{7}$$

其中:$\alpha\in\mathbb{R}^n$;$T_1,T_2\in\mathbb{R}^{n\times n}$;$t_3,t_4,t_5,t_6,t_7\in\mathbb{R}^{1\times n}$。$T_1=[\sum_{l=1}^{n}y_i y_j v_l^{\mathrm{T}}BM_i^{\mathrm{T}}M_j Bv_l]_{ij}$,$T_2=[y_i y_j w^{\mathrm{T}}BM_j^{\mathrm{T}}x_i]_{ij}$,$t_3=[4\sum_{l=1}^{n}y_i v_l^{\mathrm{T}}BVX^{\mathrm{T}}M_i Bv_l]_{1i}$,$t_4=[2\sum_{l=1}^{n}y_i v_l^{\mathrm{T}}Bwx_i^{\mathrm{T}}x_l]_{1i}$,$t_5=[2y_i w^{\mathrm{T}}BVX^{\mathrm{T}}x_i]_{1i}$,$t_6=[by_i]_{1i}$,$t_7=[1]_{1i}$,$B=(2VV^{\mathrm{T}})^{-1}$ 和 $M_i=x_i w^{\mathrm{T}}$。上式中的 $[\cdot]_{ij}$ 表示等式前的某一矩阵的第 i 行第 j 列的元素为方括号中的值,并且 v_l 表示系数矩阵 V 的第 l 列。从式(7)的形式不难看出,此问题是一个关于 α 的二次规划问题,所以可以利用一些常用二次规划的工具箱[21]得以解决。

b)为了更新超平面参数 w 和 b,需要保持 U、V 固定不变。

如此,GMNMF 的优化问题就变成了经典的软间隔 SVM 问题:

$$\arg\min_{w,b,\varepsilon}\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{n}\varepsilon_i$$

$$\text{s.t.}\quad y_i(w^{\mathrm{T}}U^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i\quad \varepsilon_i\geqslant 0,1\leqslant i\leqslant n \tag{8}$$

针对上述 SVM 问题,按照文献[22]的策略,得到对应的对偶优化问题:

$$\arg\max_{\gamma}\sum_{i=1}^{n}\gamma_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i\gamma_j y_i y_j\langle U^{\mathrm{T}}x_i,U^{\mathrm{T}}x_j\rangle$$

$$\text{s.t.}\quad 0\leqslant\gamma_i\leqslant C,i=1,\cdots,n$$

$$\sum_{i=1}^{n}\gamma_i y_i=0 \tag{9}$$

其中:$\gamma\in\mathbb{R}^n$ 是拉格朗日乘子,$\langle\cdot\rangle$ 表示两个向量的内积。为了求得具体的超平面参数值,先求解式(9)关于 γ 的二次规划问题,然后代入到式(10)。

$$w=\sum_{i=1}^{n}\gamma_i y_i U^{\mathrm{T}}x_i$$

$$b=-\frac{1}{2}\langle w,U^{\mathrm{T}}x_r+U^{\mathrm{T}}x_s\rangle \tag{10}$$

其中:$U^{\mathrm{T}}x_r$,$U^{\mathrm{T}}x_s$ 为满足 $\gamma_r,\gamma_s>0$ 以及 $y_r=-1,y_s=1$ 的任意支持向量(support vector,SV)。

c)为了更新系数矩阵 V,需要保持 U、w、b 固定不变。GMNMF 的代价函数中仅有重构误差项和图正则化项与 V 有关,因此可以得到如下优化问题:

$$\arg\min_{V}\|X-UV\|_F^2+\lambda_2 Tr(VLV^{\mathrm{T}})$$

$$\text{s.t.}\quad V\geqslant 0 \tag{11}$$

利用拉格朗日优化方法,为式(11)的约束条件引入拉格朗日乘子 $\rho_{ij}\geqslant 0(i=1,\cdots,k;j=1,\cdots,n)$,得其拉格朗日函数

$$l(V,\rho)=Tr((X-UV)(X-UV)^{\mathrm{T}})+\lambda_2 Tr(VLV^{\mathrm{T}})-Tr(\rho^{\mathrm{T}}V)$$

其中:$\rho\in\mathbb{R}^{k\times n}$。对上式的主变量 V 求偏导数并令其等于零,有 $\rho=-2U^{\mathrm{T}}X+2U^{\mathrm{T}}UV+2\lambda_2 VL^{\mathrm{T}}$。然后依据 KKT 条件 $\rho_{ij}v_{ij}=0$,可以得到 V 的更新公式:

$$V\leftarrow V\odot\frac{U^{\mathrm{T}}X+\lambda_2 VW}{U^{\mathrm{T}}UV+\lambda_2 VD} \tag{12}$$

综上所述,将 GMNMF 的求解算法归纳如下:

给定:邮件数据集$\{X,Y\}$,维数 k,正常数 λ_1、λ_2、C,收敛条件。
初始化 $V=\mathrm{rand}(k,n)$
repeat
1 若第一次迭代,$U=XV^{\mathrm{T}}(VV^{\mathrm{T}})^{-1}$;否则利用式(6)和(7)更新 U;
2 利用式(9)和(10)更新 w 和 b;
3 利用式(12)更新 V;
until 收敛
得到:基矩阵 U,系数矩阵 V,超平面参数 w 和 b。

3 仿真实验

在做实验时,按照如下步骤进行:先进行训练,即将训练样本用于各种算法得出基矩阵 U、系数矩阵 V 以及超平面参数 w 和 b;然后进行测试,将每封测试邮件的特征词条向量 x_{test} 依次投影到基矩阵 U 上,而后代入超平面的符号函数中算出相应邮件的预测类别值 $\hat{y}_{\text{test}}=\mathrm{sign}(w^{\mathrm{T}}U^{\mathrm{T}}x_{\text{test}}+b)$。

3.1 邮件数据集

实验所用到的中文邮件数据集分别来源于中国教育和科研计算机网紧急响应组(CERNET Computer Emergency Response Team,CCERT)和国际文本信息检索会议(Text Retrieval Conference,TREC)。

CCERT 中文邮件数据集(CCERT data sets of Chinese e-mails,CDSCE)共包括两个月份(六月和七月)收集的数据,其中有 45 396 封垃圾邮件和 18 314 封正常邮件。从中随机选取部分邮件进行一系列处理,主要包括中文分词、邮件表示和特征选择,最后得到所需的特征词条矩阵。实验所采用的 CDSCE 的词条矩阵大小为 2000×3900,其中有垃圾邮件和正常邮件各 1 950 封。中文垃圾邮件数据集 trec06c 是 TREC 于 2006 年中文垃圾邮件评测中给出的公开数据,其中有 21 766 封正常邮件和 42 854 封垃圾邮件。对 trec06c 进行类似于 CDSCE 的操作,最后得到用于实验的 trec06c 的词条矩阵大小为 2000×10000,其中有 6 631 封垃圾邮件和 3 369 封正常邮件。

3.2 评价指标体系

为了能够对垃圾邮件过滤模型的分类效果进行很好的评价,实验采用下列常见的性能评价指标[23]:查准率(precision)、查全率(recall)和调和平均数 F1。为了方便地对它们进行定义,假设待测试的邮件总数为 n,将它们用于构建好的垃圾邮件过滤模型,可得邮件判定结果如表 1 所示。

表 1 垃圾邮件过滤模型的判定情况分布

$n=a+b+c+d$	实际为垃圾邮件	实际为正常邮件
判定为垃圾邮件	a	b
判定为正常邮件	c	d

针对表 1 的变量定义,查准率的计算公式为 $P=\frac{a}{a+b}$,它体现了模型对垃圾邮件的检测能力,即查准率越大,正常邮件被误判为垃圾邮件的概率越小;查全率的计算公式为 $R=\frac{a}{a+c}$,它体现了模型识别垃圾邮件的能力,即查全率越大,被误判的垃圾邮件越少。可见,查准率 P 或查全率 R 越大,过滤模型就越好。然而,在某些模型中它们之间会相互影响(即一个大,而另一个小),因此实验将把 F1 作为主要的性能评价指标。F1 是查准率 P 和查全率 R 的调和平均,是它们的综合体现,计算公式为

$$\text{F1}=\frac{2\times P\times R}{P+R} \tag{13}$$

3.3 实验设计及结果分析

为了检验本文所构建的基于 GMNMF 的中文垃圾邮件过滤模型的性能,利用 CDSCE 和 trec06c 进行了两类不同的对比实验:一类是将 GMNMF 算法与其他有关 NMF 的算法(NMF + SVM、GNMF + SVM 和 MNMF)所构建的模型进行对比分析;另一类是将 GMNMF 算法与传统的特征抽取方法(ICA、PCA 和 LDA) + SVM 以及 SVM 所构建的模型进行对比分析。在上述两类实验中,所有涉及 SVM 的部分都由 LIBSVM 工具箱[24]来具体实现,并且传统的特征抽取方法的实现由 FastICA 软件包[25]和统计模式识别工具箱[26]来具体完成。

第一类共分为三组(A,B 和 C)实验。A 组实验是为了能够比较直观地看出 GMNMF 算法的模型要优于其他三种算法(NMF + SVM、GNMF + SVM 和 MNMF)的。在这组实验中,维数取 2,如此便可以将降维后的数据投影到一个二维平面上。至于其他参数,GMNMF 算法中 λ_1、λ_2、C 和最大循环次数依次取 0.01、100、2 和 4;MNMF 算法中 λ 取 GMNMF 的 λ_1;GNMF 算法中 λ 和最大循环次数依次取 GMNMF 的 λ_2 和 200;NMF 算法中最大循环次数取 200,后面三种算法中未提及而却涉及到的参数都与 GMNMF 的一样。利用 trec06c 进行 A 组实验,并且按以下规则分配数据:按照 1 ∶ 1 的比例随机组成训练样本和测试样本,同时保持在每种样本中正常邮件和垃圾邮件的

比例都为 1∶2。根据上述比例,最终用于实验的训练样本包含 1 500 封垃圾邮件和 750 封正常邮件,测试样本数目类似。将上述参数和数据用于本组实验,可以得到测试样本数据由经 NMF + SVM、GNMF + SVM、MNMF 和 GMNMF 分别降维后的投影情况分布图(图 1)。从图 1 可以看出:对于相同测试样本,GMNMF 算法所构建的模型能取得最高的 F1。不仅如此,经过 GMNMF 算法特征抽取后的数据变得明显比其他三个的数据更容易分类。如此看来,GMNMF 算法与其他三种算法相比具有一定的优越性。

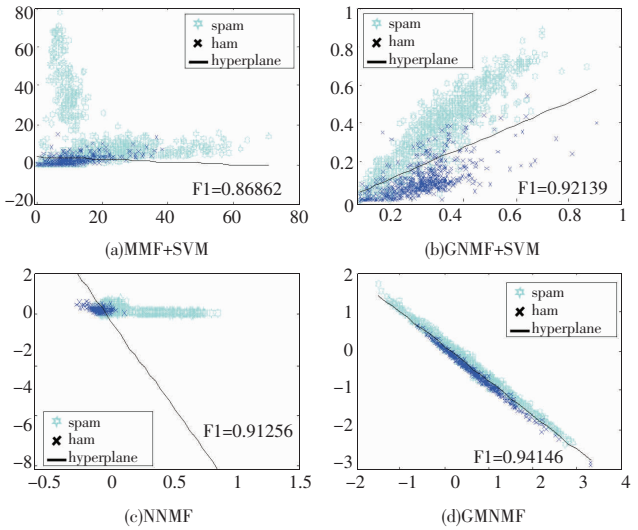

图 1　测试样本在四种算法的基矩阵下的投影分布情况

A 组实验中仅在维数 $k=2$ 进行了实验,为了验证在较小的不同维数 k 下 GMNMF 算法所构建的垃圾邮件过滤模型同样能够取得好的分类效果,进行 B 组实验。在 B 组实验中,维数 k 分别取 2、10、20、40、60 和 80,其他的参数选取和数据分配方法与 A 组中的一样。不一样的是本组将 CDSCE 和 trec06c 都用于实验,其中由CDSCE生成的训练样本有974封垃圾邮件和487封正常邮件,测试样本数目类似;由 trec06c 生成的样本数目类似于 A 组实验。将上述得到的两组样本分别用于训练和测试,可以得到 NMF + SVM、GNMF + SVM、MNMF 和 GMNMF 在不同维数下的测试分类效果如图 2 所示。从图 2 可以看出:对于每个数据集,随着维数 k 逐渐增大,四种算法的 F1 都大致呈现出逐渐增大的趋势,而且增幅总体上都逐渐减小;当 $k=10$ 时,各种算法都已经获得了较高的 F1。不仅如此,GMNMF 的 F1 在各种维数 k 上都是它们中最大的。而且绝大多数情况下,维数越小,GMNMF 的预测分类效果相对于其他三个的就越好。可见,利用经过 GMNMF 降维后的维数较小的数据进行预测分类就已经能够取得较好的分类效果。

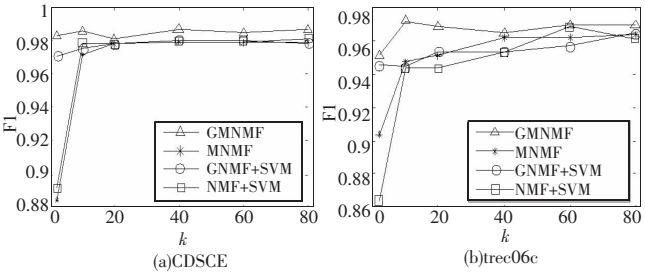

图 2　四种算法随维数 k 变化的测试分类效果

考虑到邮件数据的复杂性和随机性,在 C 组实验中,将对 B 组的每个维数 k 都分别重复运行 10 次,取得 NMF + SVM、GNMF + SVM、MNMF 和 GMNMF 所构建的模型的预测分类效果的均值和标准差。本组实验的参数取值、数据分配和数据集的选取都与 B 组实验一样,最终得到上述四种算法在不同维数 k 下重复 10 次的 F1 的均值和标准差(表 2)。从表 2 可知:针对每种算法,随着维数 k 的逐渐变大,F1 的均值大体呈现逐渐增大的趋势,可见与 B 组实验的结果相符合。不但如此,在各种维数 k 下,GMNMF 的 F1 均值都要高于其他算法的。而且,虽有少数 GMNMF 的 F1 相对于其他算法的波动较大,但整体来说,GMNMF 算法的预测分类效果还是比较好的。

表 2　不同维数下重复 10 次的 F1 的均值和标准差

k	CDSCE/%				trec06c/%			
	NMF + SVM	GNMF + SVM	MNMF	GMNMF	NMF + SVM	GNMF + SVM	MNMF	GMNMF
2	86.00 ± 0.51	95.51 ± 5.97	81.05 ± 6.91	98.73 ± 0.41	86.76 ± 0.47	93.48 ± 2.14	91.02 ± 1.65	95.18 ± 1.23
10	97.75 ± 0.35	98.01 ± 0.31	98.04 ± 0.37	98.73 ± 0.51	93.18 ± 0.99	94.86 ± 0.34	94.82 ± 0.85	96.75 ± 0.80
20	97.90 ± 0.29	98.15 ± 0.28	98.23 ± 0.27	98.76 ± 0.46	94.74 ± 0.53	94.95 ± 0.44	95.37 ± 0.49	96.53 ± 0.41
40	98.50 ± 0.35	98.13 ± 0.20	98.21 ± 0.21	98.81 ± 0.36	95.44 ± 0.72	95.55 ± 0.70	95.91 ± 0.44	96.22 ± 2.31
60	98.54 ± 0.51	98.22 ± 0.35	98.40 ± 0.39	98.78 ± 0.47	94.99 ± 1.76	95.73 ± 0.40	96.14 ± 0.38	96.77 ± 0.23
80	98.69 ± 0.34	98.27 ± 0.21	98.41 ± 0.24	98.91 ± 0.39	95.44 ± 1.88	95.91 ± 0.49	96.57 ± 0.45	96.78 ± 0.50
Avg.	96.23	97.71	95.39	98.79	93.43	95.08	94.97	96.37

第二类是将相同的测试样本分别用于 GMNMF、ICA + SVM、PCA + SVM、LDA + SVM 以及 SVM 构建的模型,对比分析它们的预测分类效果以验证本文提出的 GMNMF 算法要优于其他算法。通过第一类的 B 组实验可知,维数 $k=10$ 是一个较好的取值,所以本类实验将把 k 取为 10。其他参数的选取类似于 B 组实验,同时将仅使用 trec06c 作为样本集,分配方法同 A 组实验。将随机形成的样本集用于上面五种算法,得到它们的测试分类效果如表 3 所示。从表 3 可以看出:GMNMF 算法的 F1 要高于其他算法的,也即要优于其他各种算法。可见,本文提出的 GMNMF 算法构建的垃圾邮件过滤模型相对于其他算法的具有一定的优势。

表 3　其他算法的预测分类效果的对比　/%

算法	SVM	LDA + SVM	PCA + SVM	ICA + SVM	GMNMF
F1	96.40	81.22	94.41	82.12	96.79

4　结束语

本文提出了 GMNMF 算法,并将其用于建立中文垃圾邮件过滤模型。实验结果表明:经 GMNMF 的基矩阵投影后的数据变得比经 MNMF、GNMF 和 NMF 的基矩阵投影后要更容易分类。对于各种不同维数,GMNMF 构建的模型预测分类效果都要优于 MNMF、GNMF + SVM 和 NMF + SVM 构建的模型。不仅如此,该模型的预测分类效果要比 ICA + SVM、PCA + SVM、LDA + SVM 和 SVM 建立的模型分类效果好。在以后的进一步研究中,将推广图正则化项应用于其他有监督的机器学习算法中,以期建立更好的中文垃圾邮件过滤模型。今后,还会考虑将 GMNMF 算法应用于财务危机预测等其他方面以及设计更好的 GMNMF 的求解算法。

参考文献:

[1] YIN Hui, CHENG Feng-juan, ZHANG De-xian. Using LDA and ant colony algorithm for spam mail filtering [C] //Proc of the 2nd International Symposium on Information Science and Engineering. Washington DC: IEEE Computer Society, 2009: 368-371.

[2] WANG Ren. Feature selection strategies for spam e-mail filtering[D]. Montreal, Quebec: Concordia University, 2006: 25-31.

[3] CUI Bin, MONDAL A, SHEN Jia-lie, *et al*. On Effective e-mail classification via neural networks [C]//Proc of the 16th International Conference on Database and Expert Systems Applications. Berlin: Springer-Verlag, 2005: 85-94.

[4] DUDA R O, HART P E, STORK D G. Pattern classification[M]. 2nd ed. Hoboken: Wiley-Interscience, 2000: 114-121, 568-575.

[5] JANECEK A, GANSTERER W. E-mail classification based on NMF [C]//Proc of the 9th SIAM International Conference on Data Mining. [S. l.]: University of Vienna, 2009: 1345-1354.

[6] YOUN S, McLEOD D. A comparative study for email classification [C]//Advances and Innovations in Systems, Computing Sciences and Software Engineering. [S. l.]: Springer, 2007: 387-391.

[7] AMAYRI O, BOUGUILA N. A study of spam filtering using support vector machines [J]. Artificial Intelligence Review, 2010, 34(1): 73-108.

[8] CHANG M, YIH W, MEEK C. Partitioned logistic regression for spam filtering [C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York: ACM Press, 2008: 97-105.

[9] THURAU C, HLAVAC V. Pose primitive based human action recognition in videos or still images [C] //Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2008: 1-8.

[10] XU Wei, LIU Xin, GONG Yi-hong. Document clustering based on non-negative matrix factorization [C] //Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003: 267-273.

[11] DING C H Q, LI Tao, JORDAN M I. Convex and semi-nonnegative matrix factorizations [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(1): 45-55.

[12] KUMAR B G V, KOTSIA I, PATRAS I. Max-margin semi-NMF[C]// Proc of the British Machine Vision Conference. [S. l.]: BMVA Press, 2011: 1-11.

[13] ZAFEIRIOU S, TEFAS A, BUCIU I, *et al*. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification [J]. IEEE Trans on Neural Networks, 2006, 17(3): 683-695.

[14] CAI Deng, HE Xiao-fei, HAN Jia-wei, *et al*. Graph regularized nonnegative matrix factorization for data representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1548-1560.

[15] TRAN Q. CCERT data sets of Chinese e-mails(CDSCE) [EB/OL]. http://www.ccert.edu.cn/spam/sa/datasets.htm.

[16] CORMACK G. TREC 2006 spam track overview [C]//Proc of the 15th Text Retrieval Conference. 2006: 1-11.

[17] LIU Zun-xiong, ZHANG Xian-long, ZHENG Shu-juan. Lasso-based spam filtering with Chinese e-mails [J]. Journal of Computational Information Systems, 2012, 8(8): 3315-3322.

[18] GOLUB G H, Van LOAN C F. Matrix computations [M]. 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996: 54-57.

[19] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization [C] //Proc of Neural Information Processing Systems Conference. Cambridge: MIT Press, 2001: 556-562.

[20] LIU Hai-feng, WU Zhao-hui, LI Xue-long, *et al*. Constrained nonnegative matrix factorization for image representation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1299-1311.

[21] MATLAB optimization toolbox, user's guide, version 6.2 [K]. Natick, MA: The MathWorks, Inc., 2012.

[22] GUNN S R. Support vector machines for classification and regression, ISIS Technical Report[R]. UK: University of Southampton, School of Electronics and Computer Science, 1998.

[23] ZHANG Le, ZHU Jing-bo, YAO Tian-shun. An evaluation of statistical spam filtering techniques [J]. ACM Trans on Asian Language Information Processing, 2004, 3(4): 243-269.

[24] CHANG C, LIN C. LIBSVM: a library for support vector machines [J]. ACM Trans on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[25] The FastICA package for MATLAB[EB/OL]. (2011-02-10). http://research.ics.aalto.fi/ica/fastica/.

[26] FRANC V, HLAVAC V. Statistical pattern recognition toolbox for MATLAB[R]. Prague, Czech Republic: Center for Machine Perception, Czech Technical University, 2004.

(上接第 2621 页)

[3] TERRY L E, THOMSA D J, FERNANDA L T, *et al*. The last mile: an examination of effects of online retail delivery strategies on consumers [J]. Journal of Business Logistics, 2003, 24(2): 177-203.

[4] LEWIS M, VISHAL S, SCOTT F, *et al*. Forecasting the effects of nonlinear shipping and handling fees [J]. Journal of Marketing Research, 2003, 18(3): 221-270.

[5] LENG Ming-ming, RAFAEL B A. Joint pricing and contingent free-shipping decisions in B2C transactions[J]. Production and Operations Management, 2010, 19(4): 390-405.

[6] LEWIS M. The influence of loyalty programs and short-term promotions on customer retention [J]. Journal of Marketing Research, 2004, 41(3): 281-292.

[7] GÜMÜS M, LI Shan-ling, OH W, *et al*. Shipping fees or shipping free? A tale of two price partitioning strategies in online retailing[J]. Production and Operations Management, 2010, 21(5): 56-89.

[8] BRADEN D J, OREN S S. Nonlinear pricing to produce information [J]. Market Science, 1994, 13(3): 310-326.

[9] SARVARY M, PARKER P M. Marketing information: a competitive analysis[J]. Market Science, 1997, 16(1): 24-38.

[10] GAJANAN S, BASUROY S, BELDONA S, *et al*. Category management, product assortment, and consumer welfare[J]. Market Letter, 2007, 18(3): 135-148.

[11] TIROLE J. The theory of industrial organization[M]. Cambridge, MA: The MIT Press, 1992.

[12] CHUNG J W. Utility and production functions [M]. Cambridge: Blackwell, 1994.

[13] COTO-MiLLÁN P. Utility and production: theory and applications [M]. Berlin: Springer-Verlag, 1999.

[14] BASU A K, LAL R, SRINIVASAN V, *et al*. Salesforce compensation plans: an agency theoretic perspective [J]. Market Science, 1985, 4(4): 267-291.