由 Logistic 回归识别 Web 社区的垃圾评论

何海江,凌 云
HE Hai-jiang,LING Yun

长沙学院 计算机中心,长沙 410003
Computer Teaching Center,Changsha University,Changsha 410003,China
E-mail:haijianghe@sohu.com

HE Hai-jiang,LING Yun.Identifying comment spams of Web forums by classifier based Logistic regression.Computer Engineering and Applications,2009,45(23):140-143.

Abstract: A classifier based on Logistic Regression(LR) is employed to identify comment spams which have flooded in Web forums. Comparative study on performances of LR and Support Vector Machine(SVM) is presented.It is introduced that a relevancy coefficient vector space model named cVSM which is used to express comment archives.Some feature extractive methods are discussed,including Information Gain(IG),Mutual Information(MI),χ^2 statistic(CHI) and Document Frequency(DF).The experiments show that:The learn time of LR is less than 1/10 of SVM's.DF and IG have better performances than MI and CHI.To be compared with vector space model,cVSM has improved comment spam cognitive capability of classifier.
Key words: Logistic Regression(LR);vector space model;blog;comment spam;relevancy coefficient

摘 要:针对 Web 社区垃圾信息泛滥的问题,采用基于 Logistic 回归(LR)的分类器区分合法评论和垃圾评论,并和支持向量机(SVM)的性能对比。提出了相关度向量空间模型 cVSM 作为评论的文档表示模型,讨论了信息增益 IG、互信息 MI、χ^2 统计 CHI、文档频率 DF 等不同特征抽取方法对模型的影响。实验结果表明,LR 的训练时间不到 SVM 的 1/10;DF 和 IG 比 MI 和 CHI 表现更好;与传统的向量空间模型相比,使用 cVSM 显著提高垃圾评论识别能力。
关键词: Logistic 回归;向量空间模型;博客;垃圾评论;相关度
DOI:10.3778/j.issn.1002-8331.2009.23.039 **文章编号:**1002-8331(2009)23-0140-04 **文献标识码:**A **中图分类号:**TP391

1 前言

博客(blog)和 BBS 论坛都是互联网上的代表性应用。博客作者(也称博客,blogger)在博客上发表文章(post),记录日常生活,撰写心得体会;企业和机构也纷纷开办博客,向外发布信息,跟踪用户的反应。博客网站提供的博客工具软件,屏蔽了具体的技术细节,使用起来十分简便,又都采用免费政策,所以吸引了众多网民开设博客。博客作者往往允许来访者在文章后发表评论(comment),BBS 上热门帖子(文章)也常常吸引众多跟帖(评论)。评论成为方便和实用的交流手段,真诚的评论能激励作者发表更多更好的文章。读者也被鼓励在文章后发表评论,大多数 Web 社区可匿名随意提交评论,以此吸引人气,带动流量,还会将评论多的文章列入热门,搜索引擎也将评论作为文章评级的一项重要指标。

然而许多评论的作者并非出于交流目的,而是发表一些与文章无关的言论,或者其评论与文章有关但内容不良,将这些评论统称为垃圾评论。综合来看,垃圾评论主要有四类:(1)广告信息。包括产品推销、网站或博客推介、公司宣传等。(2)超链接评论。内容看似与合法评论语句无异,但垃圾制造者并非要引起读者注意其评论上发表的内容,而是广泛发布这些包含超链接的评论,提高其链接指向目标 Web 地址在搜索引擎的评级。(3)非商业信息。有的在评论中发表与交流话题毫无关联的信息;有的为泄私愤,在评论中散播对他人或组织不利的言论。(4)包含谩骂、下流、人身攻击等内容的语句。垃圾评论不仅占用资源、降低搜索质量,还影响读者的情绪。清除垃圾评论,保障 Web 社区的健康是一项急迫的任务。

博客和 BBS 的研究大多聚焦于文章,研究评论的报道较少,已有的讨论只限于识别第(2)类垃圾评论[1-2]。垃圾评论的识别可看作二值(合法评论、垃圾评论)文本分类问题。文本分类技术在垃圾邮件[3]、垃圾博客[4-5]、Web 文档分类等领域受到广泛研究。方法是先收集一些带标记的 Web 对象作为训练样本,再找出显著特征,最后运用贝叶斯、KNN、支持向量机(Support Vector Machine,SVM)[4-5]、Logistic 回归 (Logistic Regression,LR)[3]等算法分类。然而第(3)类垃圾评论并无显著特征,因此,提出一种相关度向量空间模型 cVSM 表示评论,再采用基于 LR 的分类算法,来识别各种类型的垃圾评论。

基金项目:长沙学院科研基金资助项目(No.CDJJ-07010110)。
作者简介:何海江(1970-),男,副教授,研究方向:Web 挖掘、数据仓库;凌云(1969-),女,高级讲师,研究方向:多媒体技术、数据库技术。
收稿日期:2008-05-29 **修回日期:**2008-08-04

2　评论的文档表示模型

向量空间模型(Vector Space Model，VSM)是最常见的文档表示模型，在 VSM 中，文档被看作以词或词权所张成的向量。但评论不单具有普通文本的特点，它还与相应文章语义关联。同样一篇评论，对文章 A 来说是合法评论，评论文章 B 时则可能变为垃圾。基于 TF*IDF，提出一种扩展 VSM 的相关度向量空间模型 cVSM，若评论文档有 n 个特征词，则定义评论为向量$(\theta,\omega_1,\omega_2,\cdots,\omega_n)$，$\theta$是评论与文章的相关度，而 ω是特征词的权重。

2.1　词权的计算

计算特征词词权 ω 的常用方法为 TF*IDF，词 w 在文档 d 的词权 $\omega_w=f_{d,w}*idf_w$，其中 $f_{d,w}$是 w 在 d 中出现的频率，idf_w为 w 的反文档频率，$idf_w=\ln(N/f_w)$。表 1 是本文使用的符号表，N、f_w等符号的意义详见表 1。

表 1　cSVM 使用的符号

q	搜索文档，对应博客文章
d	待检文档，对应评论
N	文档集包含的文档篇数
f_w	词的文档频率，文档集包含词 w 的文档篇数
f_d	文档 d 包含的词个数
$f_{d,w}$	文档 d 包含词 w 的个数

在 cVSM 中，用 IDF 代替 idf，IDF=idf/max(idf)，其中 max(idf)是所有词 idf 的最大值。为降低高频词对低频词的抑制，采用下式对词权规范化处理：

$$\omega_w=\frac{f_{d,w}*IDF_w}{\sqrt{\sum_{word\in d}(f_{d,word}*IDF_{word})^2}}$$

2.2　评论与文章的相关度

基于 TF*IDF 的内积、余弦值[6-7]等传统相关测度在搜索引擎和 Web 信息检索系统得到广泛运用。余弦值 Cosine 定义为：

$$\mathrm{Cosine}(q,d)=\frac{\sum_{w\in q\cap d}\ln(1+f_{d,w})*IDF_w*\ln(1+f_{q,w})*IDF_w}{\sqrt{\sum_{w\in q}\ln^2(1+f_{q,w})*IDF_w}*\sqrt{\sum_{w\in d}\ln^2(1+f_{d,w})*IDF_w}}$$

两文档的相关度，即语义关联程度，等于两个文档 VSM 向量的相似系数。两个文档越相似，向量的夹角越小，Cosine 值越大。

然而，这些相关度在度量 Web 社区的评论和文章时表现并不好。在垃圾评论识别中，搜索对象为长文本(文章)，待检对象为短文本(评论)，由此，在 TF*IDF 和 VSM 的基础上，提出一种更合适的相关度 CorrPC。

$$\mathrm{CorrPC}(q,d)=\frac{0.5+\sum_{w\in q\cap d}\ln(1+f_{d,w})*IDF_w}{0.5+\sum_{w\in d}\ln(1+f_{d,w})*IDF_w}$$

垃圾评论往往要发布与文章无关的内容，内容越多，也就是评论的信息熵(IDF 总和)越大，越能引起读者的注意。相对于垃圾评论来说，合法评论与文章相同的词语则更多。在 CorrPC 中，分母表示评论的信息熵，分子表示两者相关内容的信息熵。CorrPC 值越大，评论与文章的相关度也越大。去除停用词后，d 可能为空集，为避免除数为 0，在分母上加一个平滑值 0.5。另外，Web 社区大多数评论都只有寥寥几个字，而在 VSM 中只有相同词语才认为相关，语义相近，甚至是同义词都认为无关，误差无法避免。当评论与文章没有相同词时，为了使文本短的评论更倾向于被判为合法评论，在分子上也加平滑值 0.5。显然有 0<CorrPC≤1。

3　特征选择和文本分类算法

依据 VSM 或 cVSM，中文文本的特征向量往往达到数十万维，如果不经筛选而直接在全部特征上分类学习，不仅大幅度增加分类器的训练时间，还会引入一些噪声特征，降低分类性能。采用四种特征抽取方法：文档频率(Document Frequency，DF)、信息增益(Information Gain，IG)χ^2统计(CHI)、互信息(Mutual Information，MI)。CHI 度量词 w 和文档类别之间的相关程度，取两类最大 CHI 值为 w 的 CHI，即 CHI(w)=max(CHI(w，合法评论)，CHI(w，垃圾评论))。MI 的计算亦取两类的最大值。

DF+[7] 被证明在中文环境是一种有效的特征选择方法，以 DF+IG 为例，先将 DF 小于阈值的词删除，再计算剩余词的 IG，可减少噪声词对分类算法的影响。

Logistic 回归将线性判别函数转换为样本分布的后验概率对数形式，在许多领域都有成功的应用。依据统计学习理论，在 LR 中引入分类判别函数的间隔。对两类问题，给定一系列样本 $T=\{(x_1,x_1),\cdots,(x_l,y_l)\}$，其中 $\boldsymbol{x}_i\in R^n$ 是输入向量，$y_i\in\{+1,-1\}$ 是类标签，$i=1,\cdots,l$；LR 学习问题的目标为[8]：

$$\min_{w,b}\frac{1}{2}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+b^2)+C\sum_{i=1}^{l}\ln(1+e^{-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)})$$

其中 $\boldsymbol{w}^{\mathrm{T}}$ 是 $\boldsymbol{w}$ 的转置，C 是间隔和对数似然的平衡因子。为解决大规模的学习问题，Lin[8]使用信赖域的 Newton 方法求上述非约束问题的最优解。

支持向量机则使用结构风险最小化构造决策超平面，具有很好的泛化能力，在文本分类中得到广泛应用[4-5]。SVM 学习问题的目标为：

$$\min_{w,b,\xi}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{l}\xi_i$$

使得 $y_i(\boldsymbol{w}^{\mathrm{T}}\phi(x_i)+b)\geq 1-\xi_i$ 和 $\xi_i\geq 0$，$i=1,2,\cdots,l$，其中 C 为经验风险和模型复杂度的平衡因子。所有实验的核 $\phi(x)$为线性函数。

召回率 Recall、准确率 Precision、F1 是常用的分类器有效性评价指标，令被正确判为垃圾的评论数为 N_{stos}、数据集实际存在的垃圾评论数为 N_{spam}、被分类器判为垃圾的评论数为 N_{mods}，有：

$$Recall=\frac{N_{stos}}{N_{spam}}\quad Precision=\frac{N_{stos}}{N_{\mathrm{mod}\,s}}\quad F1=\frac{2*Precision*Recall}{Precision+Recall}$$

4　实验结果及分析

与报纸、杂志等媒体不同，Web 社区的文本语法并不严谨，并且有许多的网络用语，因此计算博客和 BBS 的 IDF 很有必要。编写了一个爬虫程序，从博客网站和 BBS 论坛随机下载了大量文章，涵盖教育、娱乐、科技、财经、体育等各个方面。每篇文章不少于 150 个字，筛选后得到 N=141 423 篇。使用中科院计算所汉语分词软件 ICTCLAS，将文档词频<10 的词 f_w 一律设为 3，只保留 85 601 个词及其 idf。这样做，不仅提高计算速度，

也可部分消除由于文章采样不均衡引入的噪声。不在保留词集里的词,其 idf 统一为最大值 max(idf)=ln(*N*/3)。将 f_w 超过 2/3*N 的词归类为停用词,如“的”、“是”等。所有实验都先消除文章和评论的停用词,后文不再赘述。

实验分为 5 个部分:数据集构造、参数 *C* 影响分类性能的分析、LR 与 SVM 学习时间对比、选择文本特征抽取方法、文档表示模型比较。

用 LIBLINEAR(http://www.csie.ntu.edu.tw/~cjlin/liblinear)实现的算法完成所有 LR 实验,以 LR 代称。用 LIBSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm)实现的算法完成所有 SVM 实验,以 SVM 代称。

4.1 数据集构造

构造了三个数据集,人工标注所有评论。训练数据集 BfTrain,下载了 199 篇文章(以博客文章为主),3 005 条评论,其中垃圾评论 959 条。博客测试集 BlogSet,包含 83 篇文章,2 165 条评论,其中垃圾评论 692 条。BBS 论坛测试集 BBSSet,包含 51 篇文章,523 条评论,其中垃圾评论 160 条。

4.2 *C* 影响分类性能

除 *C* 外,LR 和 SVM 均使用各自缺省的参数。相对 SVM,LR 要取较大的 *C* 才能达到 SVM 的性能。表 2 是两者分类性能 *F*1 随 *C* 变化的比较,*F*1 最大值用黑体字标示。VSM 表示用向量空间模型作为评论的文档表示模型;Cosine+(CorrPC+)表示用 cVSM 作为文档表示模型,并以 Cosine(CorrPC)值为相关度。

表 2 参数 *C* 影响 *F*1

C	1	2	4	8	16	32
SVM:VSM,MI	80.01	**80.13**	79.78	79.63	79.00	78.34
LR:VSM,MI	74.49	77.37	79.00	79.91	80.01	**80.43**
SVM:Cosine+,CHI	79.24	**80.77**	79.07	77.38	76.11	76.03
LR:Cosine+,CHI	73.25	78.77	81.87	82.20	**82.26**	81.12
SVM:CorrPC+,IG	**91.20**	90.92	89.74	88.79	88.59	87.84
LR:CorrPC+,IG	89.66	90.55	90.57	**90.58**	90.24	89.62

在 BfTrain 全部评论上分词后,共有 13 274 个特征词。分别以 5 000、6 000…12 000、13 274 个词作为评论的文本特征,在测试数据集 BlogSet 上计算平均值而得表 2 的数据。以 *MI* 为例,先计算所有词的 *MI*,再从大到小排序,当 *MI*=7 000 时,取前 7 000 个词为分类特征,最后特征词数目从 5 000~13 274 共九项逐一计算 F1 后平均。分别以 VSM、Cosine+、CorrPC+作为文档模型,再分别以 DF、IG、MI、CHI 来选择特征,在 SVM 和 LR 各得到 3×4=12 组数据,由于篇幅有限,表 2 只取其中三组数据。总体来看,当 *C*=2 时,SVM 的 *F*1 最大;当 *C*=8 时,LR 的 *F*1 最大。注意,选取高斯核函数并不能改善 SVM 的分类性能,反而大辐增加训练时间。

4.3 LR 与 SVM 学习时间对比

随着 *C* 的增大,LR 和 SVM 的优化目标值也增大,如果保持分类性能不下降,迭代次数自然增加。表 3 反映了较大的 *C* 导致较长学习时间,该结果在训练数据集上以 CorrPC+为文档模型、以 DF 为特征抽取方法而得。使用其他文档模型、其他特征抽取方法所得结论相同。由于采用了优化算法,LR 的学习时间不到 SVM 的 1/10。

SVM 和 LR 的分类性能各有千秋,与参数 *C* 有关,在不同测试数据集的表现也不尽相同;而 LR 训练时间具有显著优势,后文的实验以 LR 为主,除非特别声明,参数 *C*=8。

表 3 BlogSet 上 LR 与 SVM 学习时间比较

	5 000	6 000	7 000	8 000	9 000	10 000	11 000	12 000	13 247	平均
SVM:*C*=1	3.797	3.984	4.141	4.188	4.281	4.391	4.500	4.594	4.812	4.299
SVM:*C*=2	4.609	4.622	4.625	4.796	4.813	4.984	5.203	5.485	5.765	4.989
LR:*C*=1	0.328	0.343	0.360	0.390	0.391	0.406	0.422	0.422	0.453	0.391
LR:*C*=8	0.375	0.390	0.438	0.437	0.469	0.469	0.500	0.484	0.563	0.458

4.4 文本特征抽取方法

当训练数据集达不到一定规模时,许多词的词频较小,甚至只在垃圾评论(合法评论)出现,当这些词选入特征后,将强烈地使得分类器倾向于将评论归于垃圾评论(合法评论)。经过观察和分析,这些噪声词为数众多。要大幅度减少噪声词将是非常困难的,首先,由于中文词典的词条巨量,理想的训练数据集难于获取;其次人工标注评论类别也是非常耗时的。除 DF 外,CHI、MI、IG 都携带了类别信息,必然受到噪声词的影响;IG 平衡了类别条件概率,相比 CHI 和 MI,对噪声词的敏感程度低一些。

表 4 是使用 LR 在 BlogSet 上以 VSM 表示评论而得到的 *F*1,当以 cVSM 表示评论时,所得结果与表 4 类似。由 SVM 分类时,在 BlogSet 和 BBSSet 上 DF 都优于 IG,而由 LR 分类时,IG 略优于 DF;两个分类器在 BlogSet 上 MI>CHI,而在 BBSSet 上 CHI>MI。综合来看,DF≈IG>MI≈CHI,考虑到 DF 的时间复杂度为 O(*n*)(*n* 是文档的词个数),另三者为 O(n^2),以 DF 来抽取文本特征最合适。由于样本集以博客文章为主,在 BBSSet 上的性能都比 BlogSet 上相应项要差。

表 4 不同特征抽取方法的 *F*1 表现

	5 000	6 000	7 000	8 000	9 000	10 000	11 000	12 000	13 247	平均
DF	81.75	82.13	81.70	82.15	81.77	82.05	81.96	82.06	82.46	82.00
IG	81.92	82.25	82.16	82.65	82.10	82.14	82.05	81.99	82.46	82.19
MI	66.89	81.85	80.24	81.03	81.15	81.86	81.54	82.15	82.46	79.90
CHI	76.68	76.73	77.27	77.21	79.61	79.31	77.80	76.87	82.46	78.20

当特征词超过 5 000 时,分类性能并不随特征增多而明显改善,说明大多数特征对分类性能没有实质帮助。采用 DF+[7]的特征抽取方法,以 *DF*+*CHI*=2 000 为例,先删除所有 *DF*<3 的词,再计算剩余词的 CHI,按照顺序选择最大 CHI 值的 2 000 个词为特征。

表 5 是使用 LR 在 BBSSet 上以 Cosine+表示评论而得到的 *F*1,当以 VSM 或 CorrPC+表示评论时,所得结果与表 5 类似。综合来看,DF+IG 的性能最佳,并且较少的特征就能取得与 5 000~13 274 个特征相当的效果。在 BlogSet 上也分别以 VSM 和 cVSM 测试,所得结果与上述结论吻合。

表 5 不同 DF+方法的 *F*1 表现

	1 000	1 500	2 000	2 500	3 000	3 500	4 000	4 500	4 843	平均
DF+IG	77.7	78.10	79.57	80.85	81.82	82.52	81.56	81.85	81.00	80.55
DF+MI	63.6	72.44	75.94	76.23	77.61	78.68	79.12	80.14	81.00	76.08
DF+CHI	76.6	75.00	77.37	78.26	79.29	79.15	77.26	80.14	81.00	78.23

4.5 文档表示模型比较

图 1 是模型在 BlogSet 上的 *F*1 比较,以 DF+IG 选择特征。尽管 DF+减少了噪声词,但阈值 3 过小,无法完全消除噪声词的影响,在图中的具体表现为 *F*1 并非随特征数增多而单调增大。相关度特征的加入,召回率和 *F*1 得到大幅提高,将 cVSM

作为文档表示模型，显著提高了分类器的垃圾评论识别能力，而 CorrPC 比 Cosine 更适合度量评论和文章的相关程度。实际上，当参数 C=1 时,CorrPC+比此处的 C=8 表现更好，而 Cosine+和 VSM 则要差 4~7 个点。这在其它特征抽取方法的测试结果中也都得到证明。

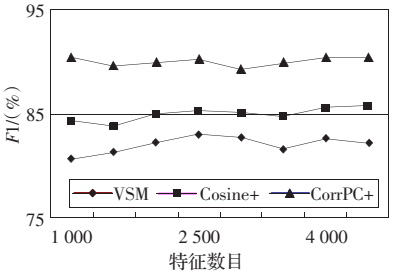

图 1　BlogSet 上 $F1$ 比较(LR,C=8,DF+IG 选择特征)

5　结束语

Web 社区的繁荣,产生了大量文本形式的文章和评论。由于网站的开放性,许多广告信息、恶意和不良信息在社区泛滥。提出一种评论表示模型 cVSM,不仅包含特征词,还考察评论和文章的相关程度,利用 LR 分类器,能有效识别垃圾评论。还讨论了 DF、MI、IG、CHI 以及 DF+等特征抽取方法的表现。以 CorrPC+表示文档模型,采用 DF+IG 抽取特征,相比其他组合形式,识别能力最强。如果使用语义相似度代替词串的匹配,性能将更优,这将是进一步的研究方向。

参考文献:

[1] Niu Yuan.A quantitative study of forum spamming using context-based analysis[C]//Proceedings of the 14th Annual Network and Distributed System Security Symposium,San Diego,CA,2007:79-92.

[2] Mishne G,Carmel D.Blocking blog spam with language model disagreement[C]//Proceedings of the 1st AIRWeb.New York:ACM,2005:1-6.

[3] 刘震,谭良,周明天.垃圾邮件分类的偏依赖特性研究[J].电子学报,2007,35(10):1870-1874.

[4] Kolari P.Detecting spam blogs:A machine learning approach[C]//Proceedings of the 21st National Conference on Artificial Intelligence.Baltimore:University of Maryland,2006:1351-1356.

[5] Lin Yu-ru.Splog detection using self-similarity analysis on blog temporal dynamics[C]//Proceedings of AIRWeb 2007.New York:ACM,2007:1-8.

[6] Brooks C H,Montanez N.Improved annotation of the blogosphere via autotagging and hierarchical clustering[C]//Proceedings of the 15th International Conference on World Wide Web.New York:ACM,2006:625-632.

[7] 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004,18(1):26-32.

[8] Lin C J,Weng R C,Keerthi S S.Trust region newton methods for large-scale logistic regression[C]//Proceedings of the 24th International Conference on Machine Learning.New York:ACM,2007:561-568.

(上接 129 页)

[3] Luck R,Ray A.Experimental verification of a delay compensation algorithm for integrated communication and control systems[J].International Journal of Control,1994,59(6):1357-1372.

[4] Zhu Qi-xin,Hu Shou-song,Hou Xia.Research of stochastic stability of networked control systems with long delay[J].Journal of Southeast University:Natural Science Edition,2003,33(3):368-371.

[5] 刘艳红,申群太.一类基于模型的时延网络控制系统的稳定性分析[J].信息与控制,2006,35(1).

[6] 关新平,黄硕,代双凤.网络控制系统的稳定性分析[J].计算机仿真,2007,24(1).

[7] 邱占芝,张庆灵,刘明.不确定时延输出反馈网络化系统保性能控制[J].控制理论与应用,2007,24(2).

[8] 崔桂梅,李小理.网络闭环系统的离散模糊控制[J].辽宁工程技术大学学报,2007,26(2).

[9] 俞力.鲁棒控制—线性矩阵不等式处理方法[M].北京:清华大学出版社,2002.

[10] Luo Xiao-yuan,Guan Xin-ping.Robust $H\infty$ control with exponent stability for time-delay uncertain systems[J].Journal of Systems Engineering and Electronics,2002,13(1):27-33.

(上接 136 页)

合模型,因而可以使分离结果更好。由于 FFT 数据块过长会削弱信号的非平稳性,因此分离效果不会继续提高,与此同时,运算量却在增加。因此实际应用中选择适当的 FFT 长度即可。

5　结论

针对频域盲源分离算法一直以来存在的“频域分离排序问题所导致的无法正确逆 STFT”的瓶颈问题,借鉴“相邻频点上分离矩阵内部结构具有高度相关性”的思想,提出了“邻频幅角比”排序算法,该算法需要额外的排序过程,但仅有少量的运算量增加,换来的是分离性能的大幅提高。仿真结果表明,“邻频幅角比”排序算法可以纠正大多数频点上的排序错误,保证逆 STFT 变换的正确进行,解决了频域盲源分离的瓶颈问题。

参考文献:

[1] Parra L,Spence C.Convolutive blind separation of non-stationary sources[J].IEEE Trans on Speech and Audio Processing,2000,8(3):320-327.

[2] Anemuller J,Kollmeier B.Amplitude modulation decorrelation for convolutive blind source separation[C]//Proceedings of the Second International Workshop on Independent Component Analysisi and Blind Signal Separation,2000:215-220.

[3] Zhou Y,Xu B.Blind source separation in frequency domain[J].Signal Processing,2000,83:2037-2046.

[4] Mitianoudis N,Davies M E.Audio source separation of convolutive mixtures[J].IEEE Trans on Speech and Audio Processing,2003,11(5):489-497.

[5] Saruwatari H,Kawamura T.Evaluation of fast-convergence algorithm for blind source separation of real convolutive mixture[C]//IEEE Pcoceeding of 6th International Conference on Signal Processing(ICSP'02),2002:346-349.

[6] 张雪峰,刘建强,冯大政.一种快速的频域盲语音分离系统[J].信号处理,2005,5(21):434-438.

[7] 虞晓,胡光锐.基于 FIR 神经网络的非线性盲信号分离[J].上海交通大学学报,1991,33(9).