SVM算法在声音广播内容监测分类中的应用研究

文 / 国家广播电视总局广播电视科学研究院　张定京　付瑞　牛泰龙　王颖　赵良福　脱羚

摘要：本文针对广播内容的智能监测，介绍了智能声音广播监测系统在广播内容识别监测方面的设计思路，以及采用SVM文本分类的方式进行广播违规内容识别和分类的实现方法。本文还介绍了对该方法和系统进行的测试和实验情况，经分析证明了此方法可以进一步提升对违规广播内容监测的效率和准确率。

关键词：广播内容监测　SVM　文本分类　机器学习　内容分类

DOI:10.16045/j.cnki.rti.2020.10.025

1 引言

近年来，调频广播由于其传播广泛、使用方便等特性，逐渐成为各种违法广播、违规广播及行业内部违章播出行为的重灾区。违法广播主要指未经过无线电管理机构和广播电视部门批准，擅自设立并利用广播的频率面向社会进行播音宣传的广播电台，主要用在违法药品、虚假广告宣传等商业活动，给社会带来了很大危害[1]。违规广播是指合法广播机构擅自更改呼号、台名、台标或者变更节目内容、播出时段，或插播其他节目等播出行为。这些非法广播，不仅扰乱了正常的无线电通讯管理秩序，对民航飞行安全、合法广播电台造成严重干扰，而且还存在危害国家安全、影响社会安定及居民身体健康的潜在风险。

对于违法广播的自动化监测，基本可以通过自动扫频，并在频域内进行比对分析的方式来做初步的甄别。目前，对于违规广播的内容监测缺乏有效的手段，往往还是通过人工监听再进行判别的方式，不仅耗费大量的人力资源，而且存在严重滞后，无法做到快速的处理。随着新一轮信息技术革命正在向智能化方向发展，提升广播监测监管的智能化和智慧化能力已迫在眉睫。鉴于上述原因，笔者及项目团队研制了一套智能声音广播监测系统，本系统利用人工智能技术，实现了对违法广播、违规广播等的自动化、智能化侦测、取证、分析。本文将重点介绍有关广播内容监测的设计思路，以及通过SVM算法对广播内容进行识别和分类的研究方法。

2 广播内容监测方法

本系统架构主要分为三部分，即监测采集前端、中心管理平台、智能语音平台。其中，监测前端的主要功能是信号监测和采集，主要包括非法频点的识别和节目内容的录制；中心管理平台是系统的中枢核心，包括设备管理、业务管理、AI识别、数据存储等功能；智能语音平台可实现语音识别、语义理解、声纹识别和语音转译等相关功能。本系统中关于广播内容识别监测的数据处理流程具体如图1所示。

内容识别监测主要是通过对合法频率广播的文字内容进行分析识别来判定。首先，定制录制规则（监测频道、时间段、录制间隔、录制时长等参数），依据录制规则对待监测广播频道的音频进行定时录制；其次，采用语音转译技术将录制的音频转译为文字文本；最后，该段文本通过预先建立的基于SVM算法的AI模型——违规广播内容分类器进行广播频道节目内容的识别分类，判断其是否存在违规内容，并识别出其违规类型（异常呼号、违规广告等）。

3 SVM算法实现文本分类

本系统依据广播节目规范性播出要求，设计制定了一套广播内容违规

检测规则，并采用基于机器学习中的SVM算法来实现对违规广播内容的分类识别。

3.1 SVM简介

支持向量机（SVM）算法，被认定是在文本分类中效率比较高的一种方法，它是一种基于统计学习发展起来的机器学习方法，其最大的特点是根据Vapnik结构风险最小化原则[2]。它的基本模型是定义在特征空间上的间隔最大的线性分类器[3]，将数据集合压缩到支持向量集合，经过学习得到分类决策函数。这种方法只需要将一定数量的文本样本通过计算表示成向量化的训练文本数据，从而提高分类的精确率，也解决了以前需要海量样本数量的问题[4]。

3.2 基于SVM的文本分类过程

SVM文本分类算法主要分四个步骤，即文本特征提取、文本特征表示、归一化处理和文本分类。

3.2.1 文本特征提取

在对文本特征进行提取时，通常采用特征独立性假设对特征选择的过程进行简化，从而达到计算时间和计算质量之间的折中。一般的方法是对文本中词汇的特征向量，通过设置特征阈值的办法来筛选最佳特征作为文本特征子集，从而建立特征模型。

特征项的提取步骤主要包括文字内容分词、统计每类词语频率、剔除停用词和单字词、统计每类词语总词频，汇总各类特征词集，合并特征词集，形成总特征词集。

3.2.2 文本特征表示

TF-IDF公式是用来计算词语的权值。在一个给定的文本中，词频（TF）具体指的是某个给定的词语在该文本出现的频率。对于在某个特定文件中的词t_i，它的重要性可以具体表示如图2中的公式（1）[5]。

逆向文件频率（IDF）是指一个词语普遍重要性的度量，某一特定词语的IDF，是由总文件数量除以包含该词语的文件的数量，再将这个商值取以10为底的对数，具体计算如图2中的公式（2）所示。其中，|D|是语料库中的文件总数，$|\{j:t_i \in d_j\}|$表示包含词语t_i的文本数量，即$n_{i,j} \neq 0$的文本数目，如果文本不包含某词语，则会出现分母为零的情况。因此，一般使用$1+|\{j:t_i \in d_j\}|$计算。

当某一个词语在特定文本中属于高频词语，并且该词语在整个文本集合中又属于低频词语，则通过图2中的公式（3）计算可以得到一个权重高的TF-IDF。

3.2.3 归一化处理

归一化就是要把所需的数据通过算法处理后，限制在一定的范围内，如图2中的公式（4）所示。其中，a为关键词的词频，min为该词在所有文本中的最小词频，max为该词在所有文本中的最大词频。由于用词频进行比较时，容易出现较大的偏差，因此通过归一化处理可以使文本分类更加的精确。

3.2.4 文本分类

经过以上三个步骤处理后，原来的文本信息已经抽象成一个向量化的样本集，然后将此样本集与训练好的模板文件进行相似度的分析计算，如果不属于该类别，则再与其他类别的模板文件进行相似度分析计算，直到确定相应的类别，这个过程就是基于SVM模型的文本分类方式。基于SVM模型的文本分类过程具体如图3所示。

本系统采用LIBSVM工具包进行

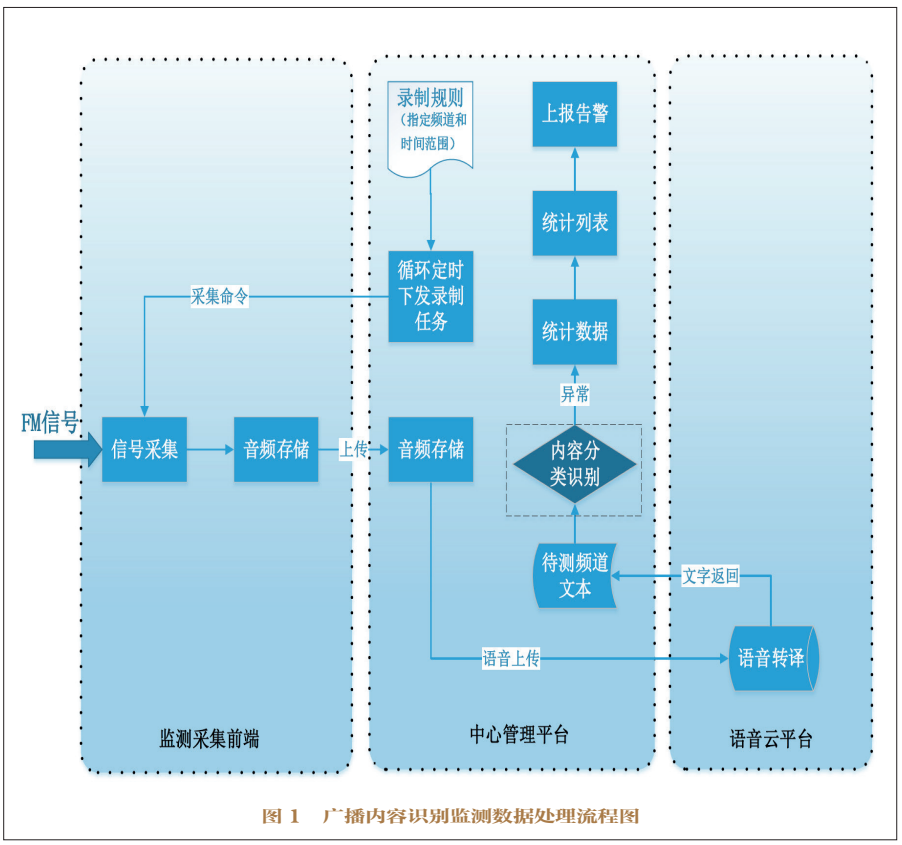

图1 广播内容识别监测数据处理流程图

SVM 的分类，实现了对广播内容文本的分类，其中被判定为违规广播内容的类型是异常呼号和违规广告。

LIBSVM 是台湾大学林智仁（Lin Chih-Jen）教授等设计开发的一个快速有效、简单易用的 SVM 模型识别与回归的软件包，该软件包提供了可在 Windows 系列系统上运行的执行文件，还提供了源代码，通过修改或改进，可以在其他操作系统上应用。该软件提供了较多的 SVM 所涉及的默认参数，所以需要调节的参数相对比较少，可以直接利用这些默认参数来解决很多的问题。该软件包还提供了交互检验（Cross Validation）的功能，可根据校验结果，选取最优的参数[2]。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$IDF_i = \lg \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (2)$$

$$TF-IDF_{i,j} = TF_{i,j} \times IDF_i \quad (3)$$

$$\frac{a-min}{max-min} = b \quad (4)$$

$$weight = 1.0 \times tf \times log_2 \frac{N}{df} \quad (5)$$

图 2

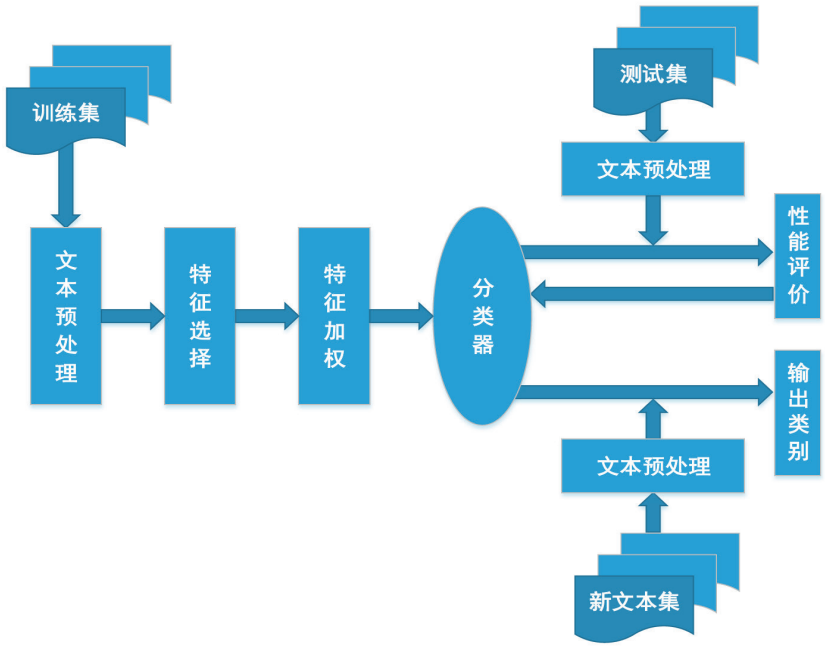

图 3　文本分类过程图

4 实验结果与分析

4.1 数据集训练与测试

4.1.1 确定训练集与测试集

训练集与测试集都属于语料库，本项目从搜狗实验室和经实际录制的广播录音并转译为文本的文件库中收集了 10 类语料库，包括交通、环境、教育、计算机、经济、军事、体育、医药、异常呼号、违规广告，后两种为录制后转译的文本，从语料库中对每个类别分别提取 50 篇，总计 500 篇作为训练集和测试集，训练集与测试集是互斥的，不相互包含。

4.1.2 对数据进行处理，生成 LIBSVM 工具需要的输入形式文件

对于训练集中的每篇文章进行分词，分词的同时过滤掉停用词。在对每篇文章进行分词、过滤停用词的同时，统计出大字典和小字典。大字典即为所有训练集的文章中出现的词汇所构成的总字典，小字典即为每个类别中该类别出现的所有文章的词所构成的字典。然后，将大字典中取出的值记为 df（文件频率），小字典取出的值记为 tf（词频）。

运用前文提到的 TF-IDF 算法，计算构建特征向量 weight，具体如图 2 中的公式（5）所示。其中，weight 表示每个词的权重，tf 即为一个类别的某个词出现的所有次数（词频），df 即为一个词在多少篇文章中出现过的次数，N 是训练集的所有文章数量。在进行测试集计算时，N 即为测试集的所有文章数量。通过对每个权值 weight 进行计算，最后生成特征向量文件。

4.1.3 利用 LIBSVM 工具进行训练及测试

在完成将训练集和测试集分别转化成所需格式的特征文件后，执行以下步骤。

（1）将训练集特征文件进行缩放，将所有没有用的数字剔除掉。

（2）将测试集所处理的特征文件进行缩放。

（3）将步骤（1）缩放后的特征文件进行训练，即得到训练模型，使用LIBSVM 软件包中的 svm-train 工具。

（4）用步骤（3）得到的模型来预测步骤（2）得到的特征文件，使用 svm-predict 工具。

经过本阶段训练测试的分类准确度可以达到 91.648%。其中，SVM 的参数 C=32，Gamma=0.0078125，Accuracy=91.648%。

4.1.4 利用交互检验方式进行参数优化

使用交互检验工具（grid.py）重新检验模型，选取最优参数，具体如图 4 所示。使用优化参数（C=128，Gamma=3.0517578125e-05）再次训练模型并进行预测，即重复上述步骤（3）和（4），最终分类的准确率可达到 94.47%（Accuracy=94.47%）。

```
[local] -1 -9 59.9078 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 11 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] -3 -9 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 9 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 3 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 15 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] -5 -9 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 7 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -7 53.9171 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -1 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -13 82.9493 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 1 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -11 86.6359 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -5 51.1521 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -15 67.2811 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 3 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 1 -9 68.6636 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 5 -3 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] -1 -3 51.6129 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 11 -3 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] -3 -3 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
[local] 9 -3 50.6912 (best c=128.0, g=3.0517578125e-05, rate=96.3134)
```

图 4　交互检验参数选取值

4.2 系统应用实验

本系统集成了上述训练验证后的 SVM 模型，应用于对广播内容违规的识别和分类。本项目对整套系统进行了现场测试，6 周时间共收到内容违规报警 946 条，852 条确认报警，94 条为误报，准确率为 90.06%。其中，广告违规报警 320 条，290 条正确报警，准确率为 90.62%；呼号报警 626 条，562 条正确报警，准确率为 89.78%。

实际应用测试结果的准确率（90.06%）比仿真预测结果的准确率（94.47%）稍低，经分析主要原因有以下两方面。

（1）本系统在录音之后的语音转译时有偏差，部分误报的广播频道因其呼号播报均为唱歌形式，无法正常转译为文字而被判断错误。

（2）训练集和测试集的样本数量较少，违规广播各类的相关词不多，影响分类精度。

后期将通过进一步扩展违规广播数据集的规模、改善训练集的标注质量，使 SVM 分类器对违规广播的分类更加准确。

5 结语

本文针对广播内容智能监测，介绍了智能声音广播监测系统在广播内容识别监测方面的设计思路，通过采用 SVM 算法进行文本分类的方式实现对广播违规内容的识别和分类，以进一步提升对违规广播监测的效率和准确率。在利用 SVM 分类器进行广播内容文本分类时，首先可以采用 TF-IDF 技术实现对广播内容文本中的关键词进行特征提取；然后再通过选取 LIBSVM 中的 Linear 核函数进行训练和参数优化，以构建用于广播内容分类的预测模型。在进行系统整体测试后的实验结果表明，内容违规报警的识别准确率在 90% 以上，略低于仿真预测结果的准确率。后期可通过进一步改善，使 SVM 分类器对违规广播的分类更加准确。

本系统可以完成声音广播监测智能判别的全流程自动化处理，基本满足试验部署的要求。该系统的应用，将有助于提高对声音广播监管的效率和时效性。RTI

参考文献：

[1] 陈爱青 . 非法调频广播自动监测系统设计与实现 [J]. 电声技术 ,2017,41(6):87-91.
[2] 沈加 . 基于 SVM 模型的新闻分类系统设计与实现 [D]. 成都：电子科技大学 ,2013.
[3] 李航 . 统计学习方法 [M]. 北京：清华大学出版社 ,2012.
[4] 孙少乙，黄志波 . 一种 SVM 多分类算法 [J]. 微型机与应用 ,2016,35(8):12-14+17.
[5] 张昭楠 . 基于 SVM 的中文文本分类系统的设计与实现 [J]. 电子设计工程 ,2016,24(16):139-141.