集成随机森林的分类模型*

邓生雄[1]，雒江涛[1,2]，刘 勇[1]，王小平[1]，杨军超[1]
(1. 重庆邮电大学 通信与信息工程学院，重庆 400065；2. 重庆市高校通信网测试技术工程研究中心，重庆 400065)

摘 要： 与集成学习相比，针对单个分类器不能获得相对较高而稳定的准确率的问题，提出一种分类模型。该模型可集成多个随机森林，并以带阈值的多数投票法作为结合方法；模型实现主要分为建立集成分类模型、实例初步预测和结合分析三个层次。MapReduce 编程方式实现的分类模型以 P2P 流量识别为例，分别与单个随机森林和集成其他算法进行对比，实验表明提出模型能获得更好的 P2P 流量识别综合分类性能，该模型也为二类型分类提供了一种可行的参考方法。

关键词： 集成学习；随机森林；带阈值的多数投票法；MapReduce；P2P 流量识别
中图分类号： TP181 **文献标志码：** A **文章编号：** 1001-3695(2015)06-1621-04
doi:10.3969/j.issn.1001-3695.2015.06.005

Classification model based on ensemble random forests

Deng Shengxiong[1], Luo Jiangtao[1,2], Liu Yong[1], Wang Xiaoping[1], Yang Junchao[1]
(1. *College of Communication & Information Engineering, Chongqing University of Posts & Telecomunications, Chongqing* 400065, *China*; 2. *Chongqing Universities Communication Network Testing Technology Engineering Research Center, Chongqing* 400065, *China*)

Abstract: Compared to ensemble learning, this paper proposed a classification model to solve the problems of relatively low and unstable accuracy in a single classifier. This model integrated multiple random forests and used majority voting method with thresholds as combination method. The implementation of this model mainly consisted of three levels, that were building the integrated classification model, the preliminary prediction of instances and combination analysis. This classification model, which had a MapReduce programming mode implementation, took P2P traffic identification as an example. This paper compared the classification model respectively with single random forests and integration of other algorithms. Finally, the experiments show that the proposed model not only has better comprehensive performance in P2P traffic identification, but also provides a viable reference method for two-class classification.

Key words: ensemble learning; random forests; majority voting with thresholds; MapReduce; P2P traffic identification

0 引言

分类是数据挖掘的一项重要任务，机器学习是实现分类行之有效的方法[1,2]。文献[3]阐述系统机器学习方法，它是关于理解和学习事物的内在机制，通过分析能对新事物作出相应预测的结合理论算法的学科。1980 年至今，机器学习以其强大的处理不同类型数据的能力和商业应用的巨大潜力在诸多领域得到应用和发展，集成学习方法是其中一个成果。集成学习的目的是将待识别实例分类为某个种类，也称做类或标签。经典的集成学习方法[4,5]已表明该方法能获得比构成它的单个分类器更高的准确率；文献[7]解释了集成方法优于单个分类器的原因。集成学习已是一个热门的课题，不同学科的研究人员都在使用集成学习方法。文献[8,9]将集成学习方法应用于面部和性别识别；文献[10]更进一步地应用该方法于情感状态识别；文献[11,12]的图像分析和室内空气污染测量运用了集成学习方法；文献[13,14]将集成学习方法运用于疾病诊断。

在网络流量识别领域，随着传统方法识别效率的降低，机器学习中单个分类器局限性的凸显，集成机器学习方法因克服传统基于端口和深度包检测方法不能很好处理的动态端口和数据加密问题，同样受到越来越多研究人员的青睐。文献[15]将决策表结合朴素贝叶斯(decision tables naive Bayes，DT-NB)、OneR 和 BP 神经网络集成来识别 P2P 流量，并评估了模型的有效性和合理性，获得了 97.27% 的平均准确率。文献[16]将朴素贝叶斯(naive Bayes，NB)、贝叶斯网络(Bayesian network，BN)和决策树(C4.5)集成，并运用 random forests 作为元分类器，对 stacking 和 voting 技术作了对比，有效地识别 P2P 流量。文献[17]将集成模型分别与 C4.5、NB 和支持向量机(support vector machine，SVM)对比，验证了模型对网络流量的分类性能。文献[18]将六个神经网络集成来识别 P2P 流量，

收稿日期：2014-04-12；修回日期：2014-06-05 基金项目：国家科技重大专项子课题资助项目(2012ZX03005002-005)；重庆市应用开发计划资助项目(cstc2013yykfA40006)；2013 年重庆高校创新团队建设计划资助项目(KJTD201312)

作者简介：邓生雄(1988-)，男，贵州贵阳人，硕士研究生，主要研究方向为数据挖掘、网络流量分类(dengsxiong@126.com)；雒江涛(1971-)，男，河南郑州人，教授，博导，主要研究方向为新一代网络技术、移动互联网数据挖掘；刘勇(1990-)，男，硕士研究生，主要研究方向为数据挖掘、网络流量分类；王小平(1988-)，男，硕士研究生，主要研究方向为数据挖掘、网络流量分类；杨军超(1988-)，男，硕士研究生，主要研究方向为数据挖掘、网络流量分类.

分别与单一 BP 神经网络、决策树、朴素贝叶斯和支持向量机算法进行比较，验证其模型的有效性。上述研究广泛地集成了朴素贝叶斯和支持向量机等算法应用于流量识别。不同的是，本文提出一种集成分类模型，该模型集成多个基于 MapReduce 的随机森林，并以带阈值的多数投票法来决策出最终预测结果。

随机森林是 Breiman[19] 于 2001 年提出的一种分类和预测算法。该算法是多棵未修剪的决策树的集合，每棵决策树的训练集源于 Bagging 方法抽样于原始训练集；在每棵树构建过程中，每个节点根据信息增益挑选最好的分裂特征；最终，每棵树同等权重投票决策出每个实例的预测类型。随机森林因其简单易行的决策机制和良好的分类性能在文本和语言的验证和处理、脸部识别和医疗等诸多方面得到应用[20~23]。随机森林的建模和预测机制是本文集成学习模型的思想来源，目的是获得较高和稳定的准确率。本文的随机森林分类算法基于 Apache Mahout(http://mahout.apache.org/) ，Mahout 旨在建立起大型的机器学习库，它包含聚类、分类和协同过滤的相应算法，并以 MapReduce 编程模式实现于 Hadoop 平台。这样的算法实现为机器学习提供了一种新的实现方式，同时也便于并行的运行方式和拓展。

借鉴随机森林多棵树投票的思想，本文提出一种分类模型，该模型首先将多个随机森林集成，然后再通过带阈值的多数投票法结合多个分类器获得最终输出。显然，在多个森林投票过程中对票数的限制能影响到最终的分类性能。

1 集成随机森林模型

1.1 集成模型

集成学习通常有两种集成框架分别为非独立和独立方法[24]。独立方法有着便于并行、提高分类器预测性能和减少运行时间的优点，本文采用了独立方法。集成随机森林与随机森林最大的不同之处在于本文通过集成多个随机森林获得预测类型和两个层次的投票机制。

本文的集成模型流程大致分为三步，总体结构如图 1 所示。

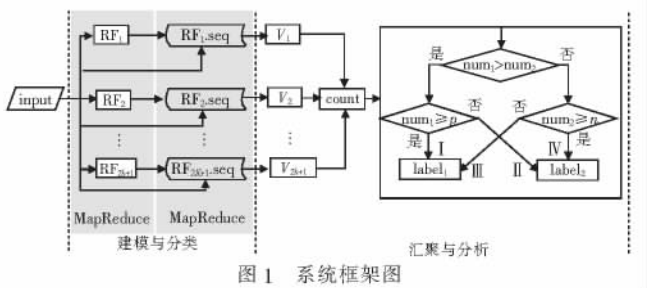

图 1 系统框架图

a) 输入，即数据集构造与输入。该步骤又分为数据预处理和中间过程，数据预处理过程需多次采集训练和测试所需的数据，然后计算相关特征，构成特征集。中间过程须先用数据预处理所构建的训练集进行必要的特征描述，同时利用数据预处理构造测试集。

b) 建模与分类。该步骤以 MapReduce 编程方式实现，在步骤 a) 的基础上首先以不同训练集训练多个分类器。文中构建的是 $2k+1$ 个基于随机森林的模型，分别为 $RF_1 \sim RF_{2k+1}$，待训练完成生成同等数目的模型文件 $RF_1.seq \sim RF_{2k+1}.seq$。再预测分类，须先加载已生成的模型文件和需预测的测试集。对于某个特定的分类器，每棵决策树对测试集的每条流作出决策并投票出流的初步应用类型，即标签(label) ，以 $V_1 \sim V_{2k+1}$ 表示每个分类器对每条流的决策。

c) 汇聚与分析。在本文的工作中，多数投票法(majority voting) 是基本的结合方法(combination method) 。在此基础上，本文以阈值限制投票过程的进行。就目前工作，本文研究的是二类型分类(two-class classification) 。在获取到同一条数据对应每个分类器预测出的标签，进行票数统计，以 num_1 和 num_2 表示多个分类器对两个类型投出的票数。为提高召回率和精确率，先判断 num_1 和 num_2 的大小，若前者大，再判断 num_1 票数是否不小于阈值 $p(k+1 \leqslant p \leqslant 2k+1)$，肯定则预测为 $label_1$，否定则预测为 $label_2$，如式(1) 所示；若后者大，再判断 num_2 是否不小于阈值 $n(k+1 \leqslant n \leqslant 2k+1)$，肯定则判定为 $label_2$，否定则判定为 $label_1$，如式(2) 所示。

$$\text{预测类型} = \begin{cases} label_1 & num_1 \geqslant p \\ label_2 & num_1 < p \end{cases} \quad (1)$$

$$\text{预测类型} = \begin{cases} label_2 & num_2 \geqslant n \\ label_1 & num_2 < n \end{cases} \quad (2)$$

与单个分类器相比，随机森林已是一种分类性能较好的分类算法，这是本文集成模型性能提升的良好保证。提升性能须从召回率和精确率两个角度考虑，引入混淆矩阵如表 1 所示。

表 1 混淆矩阵定义

classified as→	$label_1$	$label_2$
$label_1$	真正例(TP)	伪反例(FN)
$label_2$	伪正例(FP)	真反例(TN)

假定 $lable_1$ 是分类所需的类型，并且参数 p 和 n 初始值均为 $k+1$。要提高的分类性能，对两种情况分析如下：

a) 提高参数 p。该情况如图 1 中 I 所示，要求实际大部分 $label_1$ 的票数不低于 p，即能基本维持 TP 数量，尽可能少增加 FN 数量。同时又如图 1 中 Ⅲ 所示，提升该参数可阻止 $label_2$ 误识为 $label_1$，即降低 FP 数量。该参数的改变能适度提高精确率。

b) 提高参数 n。如图 1 中 Ⅱ 所示，该参数的提高能阻止实际的 $label_1$ 实例误识为 $label_2$，即增多 TP 数量降低 FN 数量。但要求实际绝大部分 $label_2$ 的票数不低于 n，即能基本维持 TN 数量，尽可能少增加 FP 数量，即图 1 中 Ⅳ。该参数的改变可以适度地提高召回率或精确率。

由上可知，参数 p 和 n 是一种调度有限的、相互制约的参数组合。尽管可信度的高低不能意味着一个预测的正确与否，但参数 p 和 n 的提高分别可增大预测为 $label_1$ 和 $label_2$ 时的可能性。然而，如果有较高的可信度，一个训练良好的模型作出的预测往往是正确的。同时，针对不同领域的数据，组合参数可能不尽相同。综合来看，上述原理可从两方面提高准确率。

1.2 模型性能评估指标

机器学习领域中，评价指标表征着模型性能优劣。本文以正确率(precision) 、召回率(recall) 和准确率(accuracy) 来评估模型的性能。正确率给出预测为正例的样本中真正正例的比例；召回率给出的是预测为正例的真实正例占所有真实正例的比例；准确率等于正确分类的样本占所有样本的比例，是召回

率和正确率的综合体现[25]。正确率、召回率和准确率公式分别如下：

$$\text{precision} = TP/(TP + FP) \tag{3}$$

$$\text{recall} = TP/(TP + FN) \tag{4}$$

$$\text{accuracy} = (TP + TN)/(TP + FN + FP + TN) \tag{5}$$

如表 1 所示，*TP* 表示真正例(true positive)，*FP* 表示伪反例(false negative)，*FP* 表示伪正例(false positive)，*TN* 表示真反例(true negative)。

1.3　特征选择

特征选择是机器学习的重要环节，选择合适的特征可以保证识别的准确率，降低特征维度可以节省建模时间。本文以 P2P 流量识别来验证集成模型的分类性能，便用五元组〈源 IP，目的 IP，源 port，目的 port，协议类型〉分流。特征选择须考虑特征与类别之间、特征与特征之间的相关性，因此本文借鉴文献[26]的特征选择算法 ReliefF 结合 CFS，该算法是基于 Weka(http://www.cs.waikato.ac.nz/ml/publications.html)机器学习工具实现。从 50 个初始特征选择出 28 个特征作为本文的特征集，分别是基于端口(源/目的)、TCP 标志(TCP 相关标志位的报文个数)、流长度特性(双向子流长度统计特征和总报文和第 1、2 个报文大小)、时间特性(流持续时间和后向子流间隔时间)和 5 报文长度区间特定的包个数和前后向包的数量等 28 个特征。

1.4　模型实现

基于 MapReduce 的随机森林实现主要分以下三个步骤：

a) 特征描述(describe)。须加载由特征集构成的多条数据，即训练集，完成描述后会生成描述文件(默认为 forest.info)。描述须严格按照数据集中特征定义的顺序和个数进行解释，如“-d”“28”“N”“L”表示数据集中有 28 个 numeric 类型特征、1 个 label 类型特征。该描述区别于*.arff 文件，它是机器学习工具 Weka 的数据输入格式。输入数据无须在文本中逐行定义特征的数据类型和相关声明。

b) 建立模型(buildForest)。模型建立需要加载训练集和步骤中生成的解释文件，并指明模型文件保存路径，训练完成会生成模型文件(默认为 forest.seq)。若须调整识别性能，可设置相关建模参数，如本文用到的参数：树的棵数(t)、节点随机选择特征数(sl)等。

c) 分类预测(testForest)。预测时需要加载预测数据和步骤中模型文件，最终输出预测应用类型。

2　测试和分析

2.1　测试平台

基于 MapReduce 的机器学习算法须在 Hadoop 平台上运行，为开发和测试，在虚拟机 Ubuntu 13.04 系统中分别使用 Hadoop 1.1.2 伪分布平台、Mahout-distribution 0.8 和主要开发软件 Eclipse-Linux 4.3.0。其中，为使 Eclipse 运行 Hadoop 程序，使用插件 Hadoop-eclipse-plugin 1.0.3。在流量采集方面，利用 Wireshark-win32 1.10.0 实时的网络抓包工具，使用单个校园网络中的主机开启 P2P 或 non-P2P 应用，多次抓取相关应用的数据包。Weka 3.7.9 用做数据集相关处理。

2.2　数据源和数据集构造

本文抓取的 P2P 数据分别有百度影音、暴风影音、PPS、PPlive、QQlive、迅雷看看和迅雷下载，均为常用的播放器和下载工具；non-P2P 数据有网页浏览、优酷视频和搜狐视频，同样也是网络流量的主要成分，如表 2 所示。

表 2　数据类型描述

P2P	non-P2P
BaiduYY，PPLive	WWW
BaofengYY，Thunder	Youku
XunleiKK，PPS，QQLive	Souhu

本文中训练集所用的数据任意大小采集于 2013 年 11 月 27 日至 2013 年 12 月 7 日期间，为保证分类器的差异性[24]，训练集(trainingset1 ~ trainingset11)使用类似于 Bagging[6] 的方法获得，不同之处在于本文的子训练集实例数小于总训练集实例数，即使用 Weka 带替换的重复随机抽样法获得训练集。测试集(testset1 ~ testset5)数据每次任意大小采集于 2013 年 12 月 31 日至 2014 年 1 月 4 日期间。训练集和测试集的流数目如表 3 所示。

表 3　训练集和测试集构成

dataset	P2P/条	non-P2P/条
trainingset1 ~ 10	7 000	10 000
trainset11	10 000	6 000
testset1	7 005	10 021
testset2	7 196	8 188
testset3	4 791	5 152
testset4	4 865	4 989
testset5	8 414	7 124

2.3　结果和分析

准确率是评估模型性能的一个重要指标，但为更好地理解模型性能，本文还计算召回率和精确率。在本文的汇聚与分析过程中，奇数个分类器易于投票过程的进行，本次实验集成 11 个随机森林，且 2.1 节中判断条件经实验得出参数 p 为 6，n 为 7，即 2.1 节中的提高参数 n。实验过程中笔者发现，同参数下，集成模型性确优于其内部的单个分类器，因此，本文把集成模型作了另外两种分类性能对比。a) 单个随机森林“树棵数 $t = 220$，节点随机选择特征数 $sl = 4$”和 11 个集成随机森林“$t = 20$，$sl = 4$”对比，文中以 trainingset11 训练的模型(single-RF)和集成模型(integrated-RFs)进行对比；b) 集成基于 Weka 的 3 个 NB[27]、4 个 C4.5[28] 和 4 个 SVM[29] 与本文集成基于 MapReduce 的 11 个随机森林对比，分别命名为 integrated-NCS 和 integrated-RFs，依次以 trainingset1 ~ 11 作为集成基于 Weka 算法的训练集和本文集成 11 个随机森林的训练集。实验 1、2 均分五次进行，即对应测试集 testset1 ~ 5。实验 1、2 测试结果如表 4 所示；实验 1 的五次测试平均结果对比如图 2 所示，实验 2 的五次测试平均结果对比如图 3 所示。

如表 4 所示，五次测试中 integrated-RFs 的召回率指标稍低于 single-RF，低出 0.2496% ~ 0.7365%。在指标精确率上，integrated-RFs 明显优于 single-RF，依次高出 2.1291%、1.4301%、1.7688%、2.6884% 和 1.0974%。尽管召回率偏低，但精确率的提高，提高了整体的准确率。如图 2 所示，integrated-RFs 的召回率、精确率和准确率均达到 96% 以上的水平，平均准确率达到 96.6695%，而 single-RF 的分类性能则呈现出相对较为参差不齐的水平。通过对比体现出集成多个随机森林

的合理性。

表 4 P2P 流量识别的测试性能

测试集	算法	recall/%	precision/%	accuracy/%
testset1	single-RF	95.657 1	94.058 1	95.458 8
	integrated-NCS	79.128 6	87.934 6	86.935 3
	integrated-RFs	95.142 9	96.187 2	96.447 1
testset2	single-RF	96.428 6	95.816 1	95.982 8
	integrated-NCS	89.730 4	93.188 1	92.128 2
	integrated-RFs	95.692 1	97.246 2	96.717 4
testset3	single-RF	96.869 1	92.708 7	94.449 4
	integrated-NCS	66.645 8	85.972	78.807 9
	integrated-RFs	96.055 1	94.477 5	95.419 5
testset4	single-RF	97.122 3	94.518 9	95.423 2
	integrated-NCS	71.942 4	94.212 7	83.965 9
	integrated-RFs	96.587 9	97.207 3	96.945 4
testset5	single-RF	98.894 7	96.263 3	96.556 8
	integrated-NCS	93.570 2	96.946 2	94.922 1
	integrated-RFs	98.654 1	97.360 7	97.818 3

如表 4 所示，integrated-RFs 的召回率在五次测试中明显优于 integrated-NCS，高出 5.0749% ~ 29.4093%。精确率方面，integrated-RFs 也优于 integrated-NCS，高出 0.4145% ~ 8.5055%。准确率方面，本文提出的集成模型明显优于集成基于 Weka 算法的模型，高出 2.8962% ~ 16.6116%，体现出集成随机森林的有效性。从图 3 可以看到，integrated-NCS 模型的低召回率是其最大的弊端，导致准确率平均结果要低出 integrated-RFs 9.3176%。数据表明，集成多个基于 Weka 的算法分类效果并不理想，初步原因有：NB 分类性能相对较差；不同算法分类机制不同，导致不同分类器对同一条流投票意见不一致，甚至分类错误；其他的汇聚与分析机制更适合集成 NB、C4.5 和 SVM；综合原因导致其不理想的分类结果。

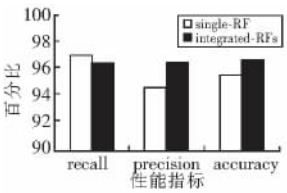

图 2 随机森林单个模型与集成模型平均性能比较

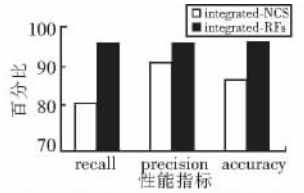

图 3 集成 Weka 算法模型和集成随机森林模型平均性能比较

准确率是整体识别性能的重要体现。在识别性能上，单个模型的预测分类不能很好地兼顾召回率和精确率，导致相对较低的准确率，体现集成模型的必要性。通过集成不同算法的比较发现，在本文的汇聚与分析机制下，集成随机森林模型性能更优，表明本文集成模型的有效性和合理性。

3 结束语

集成学习旨在以更高及更稳定的准确率将实例分类。特别地，随着互联网技术的发展，面对新时期的 P2P 应用，集成学习方法相对传统方法表现更优。本文提出集成多个随机森林的分类模型，并分别将集成模型与单个森林和集成基于 Weka 的不同算法进行性能对比。经测试表明，集成随机森林方法综合性能均优于其他两种方法，表明本文集成模型分类的有效性和合理性，但集成模型也存在性能提升的有限性。本文 P2P 流量识别获得平均准确率为 96.6695%，面对实际网络数据，还须预测性能更为良好且稳定的模型。在本文的基础上，接下来的工作是深入研究集成方法和多分类器输出的汇聚机制。在 P2P 识别方面，深入全面地分析 P2P 应用特点，建立更加完备的识别模型。

参考文献：

[1] Wang Yu, Xiang Yang, Yu Shunzheng. Internet traffic classification using machine learning: a token-based approach [C] //Proc of the 16th IEEE International Conference on Computational Science and Engineering. [S. l.]: IEEE Press, 2013: 285-289.

[2] 刘忠宝，赵文娟，师智斌. 基于分类超平面的非线性集成学习机 [J]. 计算机应用研究，2013，30(5)：1361-1364.

[3] 闫友彪，陈元琰. 机器学习的主要策略综述 [J]. 计算机应用研究，2004，21(7)：4-13.

[4] Wolpert D H. Stacked generalization [J]. Neural Networks, 1992, 5(2): 241-259.

[5] Quinlan J R. Bagging, boosting, and C4.5 [C] //Proc of the 13th National Conference on Artificial Intelligence. [S. l.]: AAAI, 1996: 725-730.

[6] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[7] Dietterich T G. Ensemble methods in machine learning [M] //Multiple Classifier Systems. Berlin: Springer, 2000: 1-15.

[8] Guo Jingming, Lin Chenchi, Chang Chehao, *et al*. Face gender recognition with halftoning-based AdaBoost classifiers [C] //Proc of IEEE International Symposium on Circuits and Systems. [S. l.]: IEEE Press, 2013: 2497-2500.

[9] Connolly J F, Granger E, Sabourin R. Dynamic multi-objective evolution of classifier ensembles for video face recognition [J]. Applied Soft Computing, 2013, 13(6): 3149-3166.

[10] Glodek M, Reuter S, Schels M, *et al*. Kalman filter based classifier fusion for affective state recognition [M] //Multiple Classifier Systems. Berlin: Springer, 2013: 85-94.

[11] Le Saux B, Sanfourche M. Robust vehicle categorization from aerial images by 3D-template matching and multiple classifier system [C] //Proc of the 7th International Symposium on Image and Signal Processing and Analysis. [S. l.]: IEEE Press, 2011: 466-470.

[12] Dang Lijun, Tian Fengchun, Zhang Lei, *et al*. A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose [J]. Sensors and Actuators A: Physical, 2014, 207: 67-74.

[13] Krawczyk B, Schaefer G. Effective multiple classifier systems for breast thermogram analysis [C] //Proc of the 21st International Conference on Pattern Recognition. [S. l.]: IEEE Press, 2012: 3345-3348.

[14] Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms [J]. Computer Methods and Programs in Biomedicine, 2011, 104(3): 443-451.

[15] 赵丹，王汝传，徐鹤. 基于集成学习的 P2P 流量识别模型 [J]. 南京邮电大学学报：自然科学版，2011，31(4)：101-105.

[16] Reddy J M, Hota C. P2P traffic classification using ensemble learning [C] //Proc of the 5th IBM Collaborative Academia Research Exchange Workshop. New York: ACM Press, 2013: 14.

[17] Dong S, Zhou D, Ding W. Traffic classification model based on integration of multiple classifiers [J]. Journal of Computational Information Systems, 2012, 8(24): 10429-10437.

[18] 徐鹤，王锁萍，王汝传，等. 基于神经网络集成的 P2P 流量识别研究 [J]. 南京邮电大学学报：自然科学版，2010，30(3)：79-83.

[19] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[20] Zhang Yang, Wang Chunheng, Xiao Baihua, *et al*. A new method for text verification based on random forests [C] //Proc of International Conference on Frontiers in Handwriting Recognition. Washington DC: IEEE Computer Society, 2012: 109-113.

(下转第 1629 页)

4　结束语

本文针对 LDA 建模结果较泛化、子话题间文本相似度较高等问题，综合考虑不同粒度的特征在表征文档时具有不同的描述能力以及传统相似度计算方法缺乏语义性等问题，提出了一种基于 LDA 模型和 HowNet 语义词典相结合的多粒度子话题划分方法 MGH-LDA。通过实验证明，相对于标准 LDA 模型直接进行子话题划分方法，该模型漏报率和误报率都有所降低，且能够很好地实现新闻子话题的划分。

今后的研究工作中，将在标准 LDA 模型建模结果的基础上，提出更多的评测指标来探究粗细粒度特征对子话题划分的影响，以便研究更多的策略实现粒度的融合来进行子话题的划分。另外，通过实验结果分析，新闻报道的体例结构特点也是影响子话题划分的重要因素，不同词性的特征词对新闻话题表达的贡献程度不同，结合新闻报道的文档结构，挖掘更具有表意性的核心特征词表示文档，可能会更进一步提高子话题的划分性能，这也是下一步的主要研究工作。

参考文献：

[1] 张阔，李涓子，吴刚，等. 基于词元再评估的新事件探测模型 [J]. 软件学报，2008，19(4)：817-828.

[2] Nallapati R，Ao F，Fu Chunpeng，*et al*. Event threading withinnews topics [C] //Proc of the 13th ACM Conference on Information and Knowledge Management. New York：ACM Press，2004：446-453.

[3] Gabriel P，Cheong F. Parameter free bursty events detection in text streams [C] //Proc of the 31st International Conference on Very Large Data Bases. New York：VLDB Endowment，2005：307-320.

[4] 于满泉，骆卫华，许洪波，等. 话题识别与跟踪中的层次化话题识别技术研究 [J]. 计算机研究与发展，2006，43(3)：489-495.

[5] 洪宇，张宇，范基礼，等. 基于子话题分治匹配的新事件检测 [J]. 计算机学报，2008，31(4)：687-695.

[6] 仲兆满，李存华，戴红伟，等. 融合内容与时间特征的中文新闻子话题聚类 [J]. 计算机科学与探索，2013，7(4)：368-377.

[7] 周学广，高飞，孙艳. 基于依存连接权 VSM 的子话题检测与跟踪方法 [J]. 通信学报，2013，34(8)：1-9.

[8] Blei D M，Ng A Y，Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research，2003，3(4)：993-1022.

[9] 李军，李涓子. 新闻专题内子话题划分 [C] //第四届全国信息检索与内容安全学术会议论文集(上). 2008：442-451.

[10] 赵爱华，刘培玉，郑燕. 基于 LDA 的新闻子话题划分方法 [J]. 小型微型计算机系统，2013，34(4)：733-737.

[11] 唐晓波，王洪艳. 基于潜在语义分析的微博主题挖掘模型研究 [J]. 图书情报工作，2012，56(24)：114-119.

[12] Blei D，Giffithe T，Jordan M. Hierarchical topic models and the nested Chinese restaurant process [C] //Advance in Neural Information Processing Systems. Berlin：Springer，2003.

[13] Zeng Jianping，Duan Jiangjiao，Wang wei，*et al*. Semantic multi-grain mixture topic model for text analysis [J]. Expert Systems with Applications，2011，38(4)：3574-3579.

[14] 张晓艳，王挺，梁晓波. LDA 模型在话题追踪中的应用 [J]. 计算机科学，2011，38(10A)：136-139.

[15] Si Xiance，Liu Zhiyuan，Li Peng，*et al*. Content-based and graph-based tag suggestion [C] //Proc of ECML/PKDD Discovery Challenge Workshop. 2009：243-260.

[16] Itawa T，Yamada T，Ueda N. Modeling social annotation data with dontent relevance using a topic model [C] //Proc of Annual Conference on Neural Information Processing Systems. 2009：835-843.

[17] Golder S A，Huberman B A. Usage patterns of collaborative tagging systems [J]. Journal of Information Science，2006，32(2)：198-208.

[18] 黄承慧，印鉴，候昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报，2011，34(5)：856-864.

[19] 刘群，李素建. 基于《知网》的词汇语义相似度的计算 [C] //第三界汉语词汇语义学研讨会论文集. 2002：59-76.

[20] 李峰，李芳. 中文词语语义相似度计算——基于《知网》2000 [J]. 中文信息学报，2007，21(3)：99-105.

[21] Papka R，James A. On-line new event detection using single-pass clustering，UM-CS-1988-021 [R]. Boston：University of Massachusetts，1998.

[22] 杨武，李阳，卢玲. 基于用户角色定位的微博热点话题检测方法 [J]. 计算机应用，2013，33(11)：3076-3079.

[23] National Institute of Standards and Technology. The 2003 topic detection and tracking task definition and evaluation plan [EB/OL]. (2012-11-21) [2013-04-26]. http//www. nist. gov/speeeh/tests/tdt/tdt2003/evalplan. htm.

[24] 张小平，周雪忠，黄厚宽，等. 一种改进的 LDA 主题模型 [J]. 北京交通大学学报，2010，34(2)：111-114.

[25] Leif A，Mark G，Keith V R. Investigating the relationship between language model perplexity and IR precision-recall measures [C] //Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2003：369-370.

(上接第 1624 页)

[21] Oparin I，Lamel L，Gauvain J. Improving Mandarin Chinese STT system with random forests language models [C] //Proc of the 7th International Symposium on Chinese Spoken Language Processing. [S. l.]：IEEE Press，2010：242-245.

[22] Huang Chen，Ding Xiaoqing，Fang Chi. Head pose estimation based on random forests for multiclass classification [C] //Proc of the 20th International Conference on Pattern Recognition. [S. l.]：IEEE Press，2010：934-937.

[23] Yang Zhiyuan，Tan Qinming. The application of random forest and morphology analysis to fault diagnosis on the chain box of ships [C] //Proc of the 3rd International Symposium on Intelligent Information Technology and Security Informatics. [S. l.]：IEEE Press，2010：315-319.

[24] Rokach L. Ensemble-based classifiers [J]. Artificial Intelligence Review，2010，33(1-2)：1-39.

[25] Witten I H，FRANK E. 数据挖掘：实用机器学习技术 [M]. 董琳，邱泉，于晓峰，等译. 2 版. 北京：机械工业出版社，2006：110-118.

[26] 刘永定，阳爱民，周序生，等. 使用机器学习算法分类 P2P 流量的方法 [J]. 计算机应用研究，2009，26(9)：3468-3471.

[27] John G H，Langley P. Estimating continuous distributions in Bayesian classifiers [C] //Proc of the 11th Conference on Uncertainty in Artificial Intelligence. San Francisco：Morgan Kaufmann Publishers Inc，1995：338-345.

[28] Salzberg S L. C4.5：programs for machine learning by J R Morgan Kaufmann Publishers，Inc，1993 [J]. Machine Learning，1994，16(3)：235-240.

[29] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery，1998，2(2)：121-167.