

计算机文本信息挖掘技术在网络安全中的应用

韩文智

(四川职业技术学院 计算机科学系, 四川 遂宁 629000)

**摘要:** 针对网络文本信息的安全性判别问题,采取改进的邻近分类算法挖掘文本.该改进邻近分类方法在传统方法定义分类特征的同时,起用共线性判别矩阵,对具有共线属性的特征合并处理.这种改进策略,不仅可以增加分类特征的准确性,也可以加快文本信息的分类进程.对 Spambase 语料库开展实验研究,从精度、召回率、联判度、误差 4 个维度对分类效果进行评价.结果显示:改进的邻近分类方法具有明显的优势,可以更加准确地区分安全文本和危险文本.

**关键词:** 文本信息;文本挖掘;文本分类;邻近分类

**中图分类号:** TP 393                      **文献标志码:** A

在信息量爆炸式增长的今天,人们生活方式发生了极大改变<sup>[1]</sup>.人们很少通过纸质文件进行信息交流,代之的是电子邮件、微博、短信、微信.这种信息交流方式确实更为便利,但也出现了新的安全隐患.部分广告人员和诈骗者,借助网络渠道向广大网络用户的邮箱、微信中发布广告信息和诈骗信息,拦截这些垃圾信息已经成为当今网络安全的重要课题之一<sup>[2]</sup>.计算机文本信息挖掘技术在信息分类、信息识别方面具有重要作用.网络信息的典型特征对于准确判断这些信息是否是垃圾信息、提升网络安全具有重要意义<sup>[3]</sup>.文献[4-10]对文本挖掘进行了研究.本文对邻近分类文本挖掘方法进行改进,提升其在网络安全中的实用效果.

1 文本挖掘和邻近分类

1.1 文本挖掘

文本挖掘是数据处理领域的一个重要分支,其操作对象主要针对文本信息.文本挖掘是从大量的文本信息中抽象、提取出具有可以理解的特征、知识,便于对文本信息进行进一步的分类、识别.

文本挖掘的过程涉及到多个环节,具体的流程如图 1 所示.文本挖掘的对象包含了各类文本信息,如期刊中的文本信息、网页中的文本信息、基于文本信息构建的数据库.文本挖掘之前,一般需要执行与处理文本信息,包括对文本信息的去噪处理、分词处理、停词处理、特征表示、特征提取.在文本挖掘这个核心阶段中,挖掘结果最终体现为文本分类、文本聚类、关联分析、趋势预测等.文中研究的重点在于文本分类.

1.2 邻近分类

邻近分类算法是文本分类的重要执行方法之一,它构建  $c$  个分类方案,并将待区分的文本分别和这  $c$  个方案进行比较,并以最接近的方案来定义文本的属性.在分类的过程中,首先要制定各个方案的描述特征,

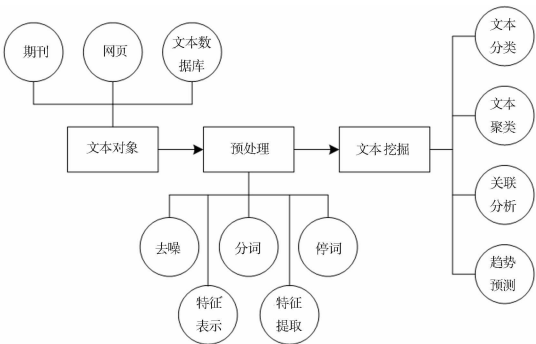


图 1 文本挖掘的流程  
Fig. 1 Process of text mining

之后,对待分类文本进行分词和特征设置,再根据相似性计算判断邻近性,其核心计算公式为

$$\rho(l_i,l_j)=\sum_{c=1}^cT_{i,c}\cdot T_{j,c}/\sqrt{(\sum_{c=1}^cT_{i,c}^2)\cdot(\sum_{c=1}^cT_{j,c}^2)}.\tag{1}$$

式(1)中: $\rho$ 表示相似性; $l_i,l_j$ 表示参照文本信息和待挖掘文本信息的特征向量; $T_{i,c},T_{j,c}$ 表示参照文本信息和待挖掘文本信息的分词.

通过式(1)可以在文本集中选取出和待挖掘文本信息相似的几个文本,判断待挖掘文本到底属于哪一个类别的公式为

$$F(l,L_j)=\sum_{l_i}\rho(l,l_i)\cdot\omega(l_i,L_j).\tag{2}$$

式(2)中: $F(l,L_j)$ 为待挖掘文本信息的最终分类结果; $\omega(l_i,L_j)$ 为待挖掘文本信息,属于某一分类权重.

## 2 邻近分类方法的改进

邻近分类方法是一类原理简单、操作方便的文本挖掘方法,但其最大的问题在于不同分类特征可能存在共线,这可能造成分类结果的不准确性.为此,在传统邻近分类方法的基础上,通过对文本特征的描述进行进一步修正.改进策略的核心思想是,将共线属性明显的文本特征进行合并,从而压缩特征向量的维度.这样,不仅能提升分类结果的准确性,也有利于算法执行速度的提高.在合并共线特征的过程中,统计变量为

$$\mathcal{R}(t_A,t_B)=\frac{S(H_1H_4-H_3H_2)^2}{(H_1+H_2)(H_1+H_3)(H_2+H_4)(H_3+H_4)}.\tag{3}$$

式(3)中: $H_1$ 为特征 $t_A$ 和特征 $t_B$ 一起出现的次数; $H_2$ 为特征 $t_A$ 出现,而特征 $t_B$ 没有出现的次数; $H_3$ 为特征 $t_A$ 没有出现,而特征 $t_B$ 出现的次数; $H_4$ 为特征 $t_A,t_B$ 都没有出现的次数.其共线性判别矩阵为

$$\mathcal{R}(t_A,t_B)=\begin{bmatrix}\mathcal{R}_{1,1}^2&\mathcal{R}_{12}^2&\mathcal{R}_{1,3}^2&\cdots&\mathcal{R}_{1,C}^2\\ \mathcal{R}_{2,1}^2&\mathcal{R}_{22}^2&\mathcal{R}_{2,3}^2&\cdots&\mathcal{R}_{2,C}^2\\ \mathcal{R}_{3,1}^2&\mathcal{R}_{32}^2&\mathcal{R}_{3,3}^2&\cdots&\mathcal{R}_{3,C}^2\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ \mathcal{R}_{S,1}^2&\mathcal{R}_{S,2}^2&\mathcal{R}_{S,3}^2&\cdots&\mathcal{R}_{S,C}^2\end{bmatrix}.\tag{4}$$

由式(4)可知: $\mathcal{R}$ 越大,特征 $t_A$ 和特征 $t_B$ 的共线特征越明显.根据这个统计变量,对传统的邻近分类方法进行改进,有如下5个操作步骤.

**步骤1** 对于邻近分类形成的各个特征计算其统计变量 $\mathcal{R}$ ,得到全部分类下全部特征的共线性判别矩阵(式(4)).

**步骤2** 在共线性判别矩阵中,对每一列元素按数值大小排列,得到其中间值 $\hat{\mathcal{R}}_j$ .如果每一列中的元素为偶数个,间值取中间两个数据的平均值.

**步骤3** 根据步骤2得到的中间值 $\hat{\mathcal{R}}_j$ ,对共线性判别矩阵中的各个元素执行归一化处理,其处理方法为 $[\bar{\mathcal{R}}_{i,j}^2]=\left[\frac{\mathcal{R}_{i,j}^2}{\hat{\mathcal{R}}_j^2}\right]$ .

**步骤4** 归一化后得到的共线性判别矩阵中,差距非常小的两个元素将被合并,从而形成更精简的特征集合.

**步骤5** 根据精简后的特征集合,采用式(1),(2)所示的方法执行邻近分类.

## 3 实验结果与分析

### 3.1 实验条件

为了验证所提出的基于改进邻近分类算法的文本挖掘方法的有效性,以网络安全检测中的应用为背景,展开实验研究.实验对象选择国际上标准的文本信息预料库 Spambase 语料库.在 Spambase 语料库中,共包含 4 600 条独立的文本信息,其中,带有危害用户信息安全的文本信息 1 800 条,其余 2 800 条为正常的文本信息.根据 Spambase 语料库的设定原则,上述 4 600 条信息可以用 58 个特征进行概括

性描述,每条文本信息到底是属于安全信息还是有危害信息,需要根据这些特征进行区分。

实验方法上,选择了传统邻近分类方法和文中方法,以便进行网络安全文本挖掘效果的横向对比。对于 Spambase 语料库中的 4 600 条文本信息,将其中 1 600 条作为训练样本,剩余 3 000 条作为实验中的检测样本。先通过 1 600 条训练样本,对两种方法进行训练,确定分类参数后,再通过另外 3 000 条文本信息检验两种方法的分类效果。

3.2 评价参数

全部文本信息的判定,只有安全信息和危险信息这两类判定结果,这是一个典型的二分类问题。为了提升判别结果的可信度,一般同时采取算法判定和专家判定两种方式。这样就出现了 4 种可能:

- 1) 算法判定结果和专家判定结果都是安全信息的文本信息,用  $T_1$  表示;
- 2) 算法判定结果为安全,专家判定结果为危险的文本信息,用  $T_2$  表示;
- 3) 算法判定结果为危险,专家判定结果为安全的文本信息,用  $T_3$  表示;
- 4) 算法判定结果和专家判定结果都是危险信息的文本信息,用  $T_4$  表示。

据此,衍生出精度、召回率、误差、联判度 4 个评价参数,其公式分别为  $p = \frac{T_1}{T_1 + T_2} \times 100\%$ ;  $r = \frac{T_1}{T_1 + T_3} \times 100\%$ ;  $e = \frac{T_2 + T_3}{T_1 + T_2 + T_3 + T_4} \times 100\%$ ;  $f = \frac{p \times r \times 2}{p + r} \times 100\%$ 。

精度、召回率、联判度都是和分类效果好坏同向的,而误差则和分类效果好坏是反向的。

3.3 实验结果

为了验证基于改进邻近分类算法的文本分类方法的有效性,设计一个网络信息安全检测分类系统软件平台。

平台以 Spambase 语料库分类的文本对象,分类方法集成了传统邻近分类方法和文中方法。软件平台的操作界面,如图 2 所示。由图 2 可知:软件平台上方为一级功能菜单区,包含了首页、用户管理、预处理、分类、趋势预测等功能,文中关注的是分类功能的设计;平台左侧是对应一级功能菜单的二级功能菜单,当前情况是选中分类菜单后其下的 3 项子功能,包括分类方法、参数评价、分类结论;平台中下方是主显示区,用于显示分类结果和对应的评价参数。

针对 Spambase 语料库的具体情况,分别选择 10 个特征进行安全信息和危险信息的区分,利用传统邻近分类方法改进邻近分类方法得到分类结果评价参数,如表 1 所示。由表 1 可知:所构建的基于改进邻近分类算法的文本分类方法,在精度、召回率、联判度、误差这 4 项评价指标上,分类效果都明显高于传统邻近分类方法;对于总数为 3 000 的测试文本信息,以 5 特征进行区分时,分类误差低于 9%。



图 2 网络信息安全检测分类系统  
Fig. 2 Network information security detection and classification system

表 1 传统邻近分类方法实验结果

Tab. 1 Experimental results of the traditional classification methods					%				
分类 C	改进前				分类 C	改进后			
	p	r	e	f		p	r	e	f
5	69.52	70.25	23.29	70.17	5	85.26	88.24	8.48	87.91
10	69.11	70.21	24.06	70.03	10	84.99	88.17	9.52	87.68
15	69.03	69.88	24.71	69.64	15	84.56	87.89	11.02	87.42
20	68.84	69.54	24.88	69.32	20	84.12	87.66	12.23	87.14
25	67.51	68.91	24.99	68.52	25	84.02	87.35	13.81	86.52
30	66.02	68.07	25.28	68.12	30	83.86	87.18	14.23	86.24
35	65.37	67.58	27.39	67.66	35	83.57	86.96	15.55	85.88
40	64.21	67.24	28.16	66.29	40	83.44	86.78	16.94	85.56
45	63.20	66.19	29.23	65.88	45	83.27	86.42	17.52	85.33
50	60.56	65.26	30.26	65.11	50	83.12	86.11	18.83	84.92

## 4 结 束 语

对邻近分类算法进行改进,并用于文本信息的安全性判别.此方法采取了共线性判别矩阵对文本信息的共线属性进行合并处理,这样可以增加属性分类的准确性,也通过合并特征属性达到提速的效果.实验结果表明,改进方法可以准确地区分安全文本和危险文本,适用于网络安全技术

### 参考文献:

- [1] DAVIES S,MOORE A. Bayesian networks for lossless dataset compression[C]// Proceeding of International Conference Knowledge Discovery and Data Mining. San Diego:ACM Press,2013:387-391.
- [2] 喻小光,陈维斌,陈荣鑫.一种数据规约的近似挖掘方法的实现[J]. 华侨大学学报(自然科学版),2008,29(3):370-374.
- [3] MERETAKIS D,WUTHRICH B. Extending naïve bayes classifiers using long item sets[C]// Proceeding of International Conference Knowledge Discovery and Data Mining. San Diego:ACM Press,2013:165-174.
- [4] ESPOSITO F,MALERBA D,SEMERARO G,et al. A comparative analysis of methods for pruning decision trees [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,19(5):476-491.
- [5] LAM S L Y,LEE D L. Feature reduction for neural network based text categorization[C]//Digital Symposium Collection of 6th International Conference on Database System for Advanced Application. [S. l. ]:IEEE Press,2015:1121-1130.
- [6] CESTNIK B,BRATKO I. On estimating probabilities in tree pruning, machine learning: EWSL-91[C]// Kodratoff Lecture Notes in Artificial Intelligence. Berlin:Springer,2015:138-150.
- [7] ANDROUTSOPOULOS G,PALIOURAS V,KARKALETSIS G,et al. Learning to filter spam e-mail: A comparison of a naïve Bayesian and a memory based approach[C]// Proceedings of 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. London:Jerry Press,2000:1-13.
- [8] SUN Lihua,ZHANG Jidong,LI Jingmei. An improved knearest neighbor system and its application to text classification[J]. Applied Science and Technology,2002,29(2):25-27.
- [9] 寸待杰,刘韶涛.采用内容挖掘的缅甸文字相似性文档检索[J]. 华侨大学学报(自然科学版),2013,34(5):521-524.
- [10] RASTOGI R,SHIM K. Public: A decision tree that integrates building and pruning[C]// Proceeding of 24th International Conference on Very Large Data Bases. New York:[s. n. ],2014:404-415.

# Application of Computer Text Information Mining Technology in Network Security

HAN Wenzhi

(Department of Computer Science, Sichuan Vocational and Technical College, Suining 629000, China)

**Abstract:** In view of the security problem of network text information, we adopt an improved neighbor classification algorithm to carry out text mining. In improved nearest neighbor method, definition and classification are carried out by traditional method, and characteristics are merged by reinstating co-linear discriminant matrix of collinear attribute features. This improved strategy not only increase the accuracy of classification features, but also speed up the classification process of text information. An experimental study is carried out on the Spambase corpus, and the classification results are evaluated from 4 dimensions. Namely accuracy, recall rate, the degree of error, and the error. Results show that the improved method has obvious advantages, and that is more accurate in the area of security text and dangerous text.

**Keywords:** text information; text mining; text classification; neighbor classification

(责任编辑: 陈志贤      英文审校: 吴逢铁)