基于关联矩阵的短信自动分类

李 锋 万小强
(东华大学计算机科学与技术学院 上海 200000)

摘 要 短信自动分类是短文本研究的热点问题。针对此问题,提出了关联强度和关联矩阵特征提取方法,并设计了基于关联矩阵的全监督学习算法。为了实现系统的自我学习,探讨了基于关联矩阵的半监督学习算法,其结合了人工矫正的主动学习算法。最后通过实例验证说明了算法的有效性。

关键词 短文本,短信自动分类,关联矩阵,半监督学习,主动学习
中图法分类号 TP391 **文献标识码** A

SMS Automatic Classification Based on Relational Matrix

LI Feng WAN Xiao-qiang
(College of Computer Science and Technology, Donghua University, Shanghai 200000, China)

Abstract SMS automatic classification is a hot issue of short text study. In this problem, this paper put forward to the feature extraction method of relational strength and the relational matrix, and designed a fully supervised learning algorithm based on relational matrix. In order to implement the system of self learning, this paper also discussed a semi-supervised learning algorithm based on relational matrix, which combines with active learning algorithm of the artificial modification. Finally the experiment results illustrate the effectiveness and efficiency of this algorithm.

Keywords Short text, SMS automatic classification, Relational matrix, Semi-supervised learning, Active learning

如今人们经常通过短信来传递相关的信息,但是广告垃圾消息、业务亲友信息的混杂也给人们带来了许多不便。短信自动分类可以有效帮助人们尽快获得所需信息。目前已有很多短信自动分类的相关研究。杨柳等人[1]提出了一种新的智能短信分类算法——朴素贝叶斯算法;李继刚等人[2]提出了结合贝叶斯和 Bigram 分词算法来实现短信自动分类的方法;王文霞等人[3]详细研究了短信文本分类的相关技术;张永军和刘金岭提出了计算词分类权重的方法[4],归纳了分类能量空间的概念[5];Chen T 和 Kan M Y[6]提出了短信的一些特征以及性质;Forman G 等人[7]提出了基于特征选择的短信分类方法;王文霞[8]提出了基于贝叶斯文本分类算法的垃圾短信过滤系统分类;张杰等人[9]提出了归一化词频贝叶斯模型的文本分类方法;董红斌等人[10]提出了一种基于关联信息熵度量的特征选择方法。

目前,主要的研究难点在于训练集的缺乏给主动学习带来的问题,以及半监督学习和主动学习相结合出现的其他相关问题[11,13]。

针对目前中文短信自动分类所面临的挑战,本文提出了一种基于关联矩阵全监督学习、半监督学习和主动学习相结合的短信自动分类算法。该算法首先利用一部分已标注短信进行训练,为了解决短信语料库缺乏的问题,本文在此基础上又引入了半监督学习。但是在实验过程中发现半监督学习存在初始学习错误问题,因此在此基础上引入了主动学习的思路来解决该问题。为了验证该算法的可行性,本文进行了三分类及多分类实验。三分类的目的是验证算法是否适合短信自动分类;多分类的目的是验证算法的健壮性,验证其是否只能在类别较少的情况下才会有较好的准确率及召回率。实验结果证明,该算法在短信自动分类上具有较好的适用性及健壮性。

1 短信的特征值提取

在短信分类中,短信中的词语是一个重要的分类特征,选择合理的特征词提取方法对短信自动分类是非常重要的。目前比较常用的特征提取方法主要包括文档频率和卡方统计[12-14]。为了说明的方便,如不特别标注,本文中所提特征词均不包含停用词[15-16]。

1.1 文档频率特征值提取

文档频率[17]主要指的是某一特征词在训练文档集中出现过的文档数。这种方式实现简单,在短文本应用中也取得了一些成果,但是它仅仅只是考虑到了特征量词频问题,忽略了特征词和类别之间的联系。

1.2 卡方统计特征值提取

卡方统计是一种假设检验方法。在文本分类中,它可以用来检验特征词 t 和文档类别 c_j 之间的关联强度。卡方统计的公式如下[18]:

$$x^2(t,c_j)=\frac{N*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (1)$$

其中,N 为训练文本集中的样本总数,A 为训练文本集中特征项 t 和类 c_j 同时出现的次数,B 为特征项 t 出现而类 c_j 不出现的次数,C 为特征项 t 不出现而类 c_j 出现的次数,D 为特征项 t 和类 c_j 都不出现的次数。研究表明,卡方统计是一种很高效的特征提取方法[19]。文献[20]表明,在 RCV1 语料库

本文受上海市自然基金项目(16ZR1401100)资助。

李 锋 博士,教授,主要研究方向为嵌入式技术、软件开发、机器学习,E-mail:lifeng@dhu.edu.cn;万小强 硕士生,主要研究方向为机器学习、软件开发,E-mail:13917251667@163.com。

中,基于卡方统计方法的性能相对较好。

1.3　关联强度的特征提取

综合上述两种方式,结合特征词词频以及特征词和类别之间的关系,本文提出了一种基于"关联强度"的特征提取方式。本文用 $AS(t,c)$ 表征特征词 t 与文档类别 c 的关联强度。以下 4 个因素决定了关联强度的大小:

(1)特征词 t 在一个类别 c 的文档中出现的次数即 $Co(t,c)$。$Co(t,c)$ 的值越大,则该特征词与该类别文档的关联性越强,因此 $AS(t,c) \propto Co(t,c)$。

(2)类别 c 中特征词的总数即 $Terms(c)$。$Terms(c)$ 越大,则某个特征词对于该类别的重要程度就越小,关联强度也就越小,因此 $AS(t,c) \propto 1/Terms(c)$。

(3)训练集中类别数即 Cf。Cf 越大,则该特征词对于某个类别的重要性也就越小,因此 $AS(t,c) \propto 1/Cf$。

(4)在保证每个类别每个文档中内容相同的条件下,文档的数目即 $D(c)$。$D(c)$ 越大,该特征词对于该类别的重要性则越低,因此 $AS(t,c) \propto 1/D(c)$。

综上所述,得到一个特征词 t 与一个类别 c 的关联强度 $AS(t,c)$ 的计算公式:

$$AS(t,c)=Co(t,c)*\frac{1}{Terms(c)}*\frac{1}{Cf}*\frac{1}{D(c)} \tag{2}$$

利用反正切归一化的方式得到最终的关联强度:

$$AS(t,c)=\frac{\arctan(Co(t,c)*\frac{1}{Terms(c)}*\frac{1}{Cf}*\frac{1}{D(c)})*2}{\pi} \tag{3}$$

其中,π 为常数 3.1415。

2　关联矩阵短信自动分类算法

机器学习[21-22]可分为无监督学习、全监督学习以及半监督学习。全监督学习主要考虑标注的样本;半监督学习主要考虑通过在未标注的样本的分类过程中实现机器的自我学习,在无人监督的情况下完成分类参数的调整。

2.1　全监督学习的关联矩阵分类方法

2.1.1　关联矩阵的计算

全监督学习就是利用已经标示出类别的短信库,通过式(3)计算特征词 t_i 和类别 c_j 的关联强度 $AS(t_i,c_j)$,从而形成特征词和短信类别关联矩阵:

$$M=(AS(t_i,c_j))_{m\times n} \tag{4}$$

关联矩阵如表 1 所列。

表 1　关联矩阵

特征词 \ 类别	c_1	c_2	c_3	…	c_n
$T(1)$	$AS(t_1,c_1)$	$AS(t_1,c_2)$	$AS(t_1,c_3)$	…	$AS(t_1,c_n)$
$T(2)$	$AS(t_2,c_1)$	$AS(t_2,c_2)$	$AS(t_2,c_3)$	…	$AS(t_2,c_n)$
$T(3)$	$AS(t_3,c_1)$	$AS(t_3,c_2)$	$AS(t_3,c_3)$	…	$AS(t_3,c_n)$
…	…	…	…	…	…
$T(m)$	$AS(t_m,c_1)$	$AS(t_m,c_2)$	$AS(t_m,c_3)$	…	$AS(t_m,c_n)$

关联矩阵 M 主要是用矩阵形式来表示多个特征词和不同短信类别之间的关系,其具体形成算法如算法 1 所示。

算法 1　完全监督学习的关联矩阵生成算法

输入:已标示短信类别的短信库 K

输出:关联矩阵 M

1. 判断短信库是否为空,不为空则执行步骤 2－步骤 3,为空则退出程序。
2. 取出 K 中的短信 S,其类别记为 c_j。
3. $D(c_j)=D(c_j)+1$。
4. 通过分词算法找出 S 中所有特征词 t_i 和出现的频度 L_i。
5. 对每个 t_i 分别执行步骤 6－步骤 8。
6. $Co(t_i,c_j)=Co(t_i,c_j)+L_i$。
7. $Term(c_j)=Term(c_j)+L_i$。
8. 针对所有的特征词 t_i 和类别 c_j,利用式(3)计算 AS(t,c),从而形成关联矩阵 M。

2.1.2　短信分类

在得到关联矩阵 M 后,针对待分类短信 S_x,其分类过程如下。

对 S_x 进行分词,形成特征词频向量:

$$T_x=[Co(t_1,S_x),Co(t_2,S_x),\cdots,Co(t_i,S_x)]^{\mathrm{T}}$$

其中,$Co(t_i,S_x)$ 表示矩阵 M 中的特征词 t_i 在 S_x 中出现的频度。

计算 MT:

$$MT=T_x*M \tag{5}$$

可知,MT 为行 $1*N$ 的行向量:

$$MT=[ASSum(S_x,c_1),ASSum(S_x,c_1),\cdots,ASSum(S_x,c_i),\cdots]$$

其中,$ASSum(S_x,c_i)$ 为短信 S_x 和类别 c_i 的"关联总值":

$$ASSum(S_x,c_i)=\sum_{j=1}^{k}Co(t_j,S_x)*AS(t_j,c_i) \tag{6}$$

MT 最大的值所对应的短信类别即为分类结果,即待测样本 S_x 所属最终类别为

$$Max(MT) \tag{7}$$

所对应的类别 c_j。

由于关联矩阵 M 是一个稀疏矩阵,因此可以利用哈希算法来存储。

为了降低矩阵的维度,可以依据关联强度,采用式(8)进行降维。

$$Max\_AS(t)=Max(AS(t,c_i)) \tag{8}$$

最后选取 $Max\_AS(t)$ 最大的前 k 个特征词作为整个训练集的特征词集合 $T(t_1,t_2,t_3,\cdots,t_k)$ 即可。

2.2　基于半监督的关联矩阵分类算法

由于标示短信库的缺乏,导致不同用户对短信分类的标准也有所不同,例如股票信息短信对某些人来讲是垃圾短信,但对某些用户可能是业务短信,因此利用在线自主学习提高短信自动分类的智能性是非常重要的。这里引入半监督学习[23],前面已经介绍了关于半监督学习的概念,这里主要讨论关于半监督学习算法的具体实现。

半监督学习过程:基于关联矩阵的半监督学习就是在具体分类过程中通过相应的算法不断完善 2.1 节中的关联矩阵 M,从而提高分类的精确度。

算法 2　半监督学习中关联矩阵修订算法

输入:待分类短信 S_x,关联矩阵 M

输出:S_x 的类别 c_j,修订后的关联矩阵 M

1. 通过分词算法找出 S_x 中的所有特征词 t_i 和出现的频度 L_i。
2. 利用 2.1.2 节方法对 S_x 进行分类,得到短信类别 c_j。
3. 对每个 t_i 分别执行步骤 4－步骤 6。
4. $Co(t_i,c_j)=Co(t_i,c_j)+L_i$。
5. $Term(c_j)=Term(c_j)+L_i$。
6. 利用式(3)分别对所有的特征词 t_i 和类别 c_j 计算 AS(t,c),从而修订关联矩阵 M。

2.3　基于半监督结合主动学习的分类

基于半监督关联矩阵的分类算法在学习上达到了自我学

习的能力，但是同时也存在一个问题，即如果在全监督学习中出现分类错误问题，那么在以后的学习中便会一直朝这个错误的方向走下去。因此，为了避免这个问题导致算法分类准确性下降，本文引入了主动学习，同时也设置了惩罚机制来提高算法的有效性。

主动学习就是利用人工标示方法对自动分类错误进行纠正，并在这个纠正过程中完成对关联矩阵的再修订。

惩罚机制的本质就是在学习过程中对与自动分类错误相关的关联强度进行惩罚性降低。

主动学习的具体实现如算法 3 所示。

算法 3 半监督结合主动学习的关联矩阵修订算法

输入：待分类短信 S_x，关联矩阵 M，自动分类的结果 c_x，手工分类结果 $c_y(c_x \neq c_y)$

输出：修订后的关联矩阵 M

1. 对每个 t_i 分别执行步骤 2－步骤 5。
2. $AS(t_i,c_x)=\alpha * AS(t_i,c_x)$，其中 α 为惩罚因子，$0<\alpha<1$。
3. $Co(t_i,c_y)=Co(t_i,c_y)+L_i$。
4. $Term(c_j)=Term(c_j)+L_i$。
5. 利用式(3)分别对所有的特征词 t_i 和类别 c_j 计算 $AS(t,c)$，从而修订关联矩阵 M。

3 实验和分析

本文通过实验来对采用关联强度特征的半监督关联矩阵分类算法以及采用关联强度特征的半监督关联矩阵结合主动学习分类算法进行评估，实验还对其特征提取方式的选取做了相应的分析[24-25]。

首先对短信类别为 3 的情况进行了实验测试，然后再对多分类进行实验测试。

实验平台为 android 智能手机，实验的中文分词采用 IK 分词算法[26]。

3.1 算法评价方法

对于短信分类的评价方法，本文采用传统的准确率(P)、召回率(R)以及 F 值来评价其优劣。对于类别 c_i，设系统分到 c_i 中的样本数目为 A，其中真实属于类别 c_i 的样本数目为 B，整个测试样本中真实属于 c_i 的样本数目为 D，则存在以下定义[27-28]：

$$P(c_i)=\frac{B}{A} \tag{9}$$

$$R(c_i)=\frac{B}{D} \tag{10}$$

$$F=P(c_i)*R(c_i)*2/(P(c_i)+R(c_i)) \tag{11}$$

3.2 测试 1

由于日常生活中祝福类、垃圾类、普通类 3 类短信的分类信息相对比较明确，因此本实验也以这 3 种类别进行测试。其中祝福类以及垃圾类是来自互联网[29]，普通类则是采用了 NUS SMS Corpus[1] 作为本次的实验数据，分别随机抽取 80%作为训练集，其余 20%作为测试集。

训练集以及测试集的数量如表 2 所列。

表 2 3 种类别测试的训练集以及测试集数量

	祝福类	普通类	垃圾类
训练集	1600	1600	1600
测试集	400	400	400

实验采用关联强度特征的全监督关联矩阵，文档频率法[19]和卡方统计法[30]的实验结果如表 3－表 5 和图 1－图 3 所示。

表 3 采用关联强度特征的全监督关联矩阵

	祝福类	普通类	垃圾类	综合
准确率(P)	0.745	0.820	0.812	0.792
召回率(R)	0.842	0.791	0.865	0.832
F 值	0.790	0.805	0.837	0.810

表 4 采用文档频率法

	祝福类	普通类	垃圾类	综合
准确率	0.725	0.810	0.803	0.779
召回率	0.821	0.775	0.824	0.807
F 值	0.770	0.792	0.813	0.791

表 5 采用卡方统计法

	祝福类	普通类	垃圾类	综合
准确率	0.735	0.811	0.798	0.781
召回率	0.812	0.768	0.846	0.808
F 值	0.771	0.788	0.814	0.791

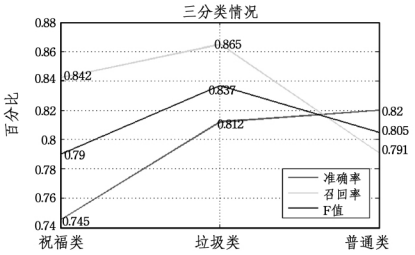

图 1 采用关联强度特征的全监督关联矩阵的分类情况

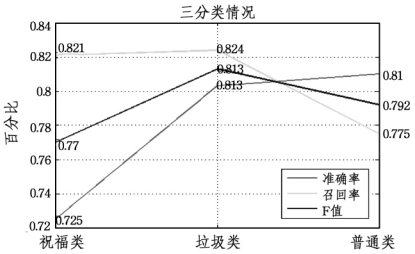

图 2 采用文档频率法的分类情况

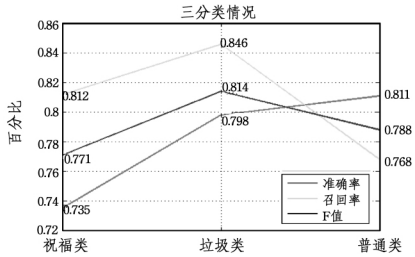

图 3 采用卡方统计法的分类情况

通过表 3－表 5 和图 1－图 3 可以得到，关联强度特征的全监督关联矩阵的分类效果优于文档频率法和卡方统计法。同时在后续会遇到大量未标注的短信，因此本文在此基础上引入了半监督学习的方法。实验选取 400 条未标注的短信，用前面全监督训练得到的矩阵对其进行分类，并在分类中通过算法 2 重新修订了关联矩阵 M 并得到了新的分类器，然后利用表 4 中的测试集对修正后的分类器进行了验证，实验结果如表 6 和图 4 所示。

表 6 采用关联强度特征的半监督关联矩阵

	祝福类	普通类	垃圾类	综合
准确率	0.821	0.909	0.982	0.904
召回率	0.962	1.000	0.820	0.927
F 值	0.885	0.952	0.893	0.910

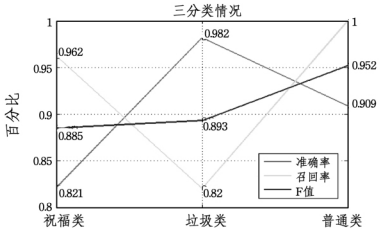

图 4　采用关联强度特征的半监督关联矩阵的分类情况

通过表 6 和图 4 可以得到，关联强度特征的半监督关联矩阵的综合准确率为 0.904，综合召回率为 0.927，综合 F 值为 0.910。由此看出，半监督学习在全监督学习的基础上提高了分类的效果。鉴于前面也提到基于半监督学习可能会存在一开始学习错误的问题，本文结合了主动学习的方法进行实验。

实验又选取了 200 条短信让系统在人工监督的情况下进行自动分类，对分类错误的短信进行人工调整，再通过算法 3 重新修订关联矩阵 M，从而得到新的分类器；然后利用表 4 中的测试集对这个修正后的分类器进行了验证，实验结果如表 7 和图 5 所示。

表 7　采用关联强度特征的半监督关联矩阵结合主动学习

	祝福类	普通类	垃圾类	综合
准确率	0.961	0.987	0.924	0.957
召回率	0.980	0.912	1.000	0.963
F 值	0.970	0.948	0.960	0.959

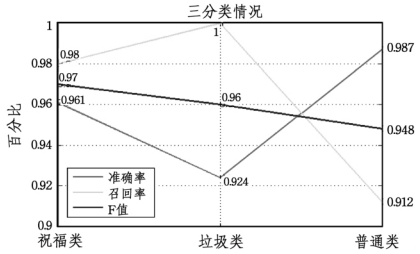

图 5　采用关联强度特征的半监督关联矩阵结合主动学习的分类情况

通过表 7 和图 5 可以得到，关联强度特征的关联矩阵结合主动学习的综合准确率为 0.957，综合召回率为 0.963，综合 F 值为 0.959。由此可以得到，结合主动学习的半监督关联矩阵的算法对其分类的准确率、召回率、F 值都得到提高。

为了验证算法的可行性，本文也将所提算法和前人研究的相关算法进行了比较，如表 8 和图 6 所示。

表 8　算法比较

	关联矩阵	朴素贝叶斯	神经网络	KNN	SVM
准确率	0.957	0.812	0.857	0.865	0.895
召回率	0.963	0.825	0.869	0.873	0.879
F 值	0.959	0.818	0.863	0.868	0.886

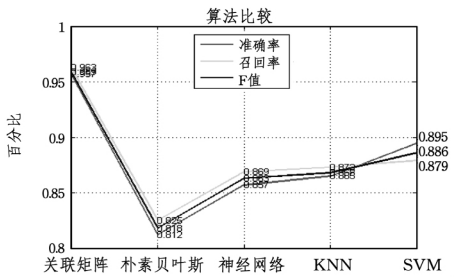

图 6　算法比较

3.3　测试 2

为了验证算法的健壮性，又选取了 4 种类别的短信进行实验验证。通过人工收集互联网上的相关类别信息，把短信分为：友情类、爱情类、健康类、节日类 4 类，再进行新的实验。同样，本文选取 80% 作为训练集，其余 20% 作为测试集，如表 9 所列。

表 9　多类别测试的训练集以及测试集的数量

	友情类	爱情类	健康类	节日类
训练集	1600	1600	800	1600
测试集	400	400	200	400

实验利用算法 1 即关联强度特征的全监督关联矩阵得到原始分类器，然后利用该分类器对表 9 中的测试集进行了实验，实验结果如表 10 和图 7 所示。

表 10　采用关联强度特征的全监督关联矩阵

	友情类	爱情类	健康类	节日类	综合
准确率	0.831	0.842	0.836	0.858	0.842
召回率	0.853	0.857	0.842	0.861	0.853
F 值	0.841	0.849	0.838	0.859	0.847

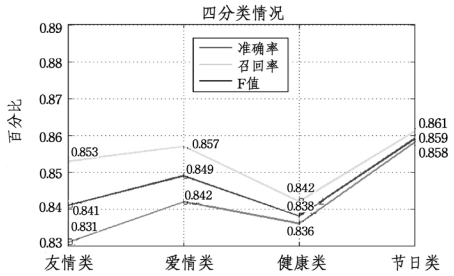

图 7　采用关联强度特征的全监督关联矩阵的分类情况

通过表 10 和图 7 可以得到，关联强度特征的全监督关联矩阵的综合准确率为 0.842，综合召回率为 0.853，综合 F 值为 0.847。

实验选取了 400 条未标注短信，然后又通过 400 条未标注短信进行实际分类，利用算法 2 重新修订了关联矩阵 M，得到新的分类器；最后通过表 9 中的测试集对该分类器进行了实验。实验结果如表 11 和图 8 所示。

表 11　采用关联强度特征的半监督关联矩阵

	友情类	爱情类	健康类	节日类	综合
准确率	0.875	0.883	0.819	0.894	0.867
召回率	0.874	0.892	0.837	0.875	0.869
F 值	0.874	0.887	0.827	0.884	0.868

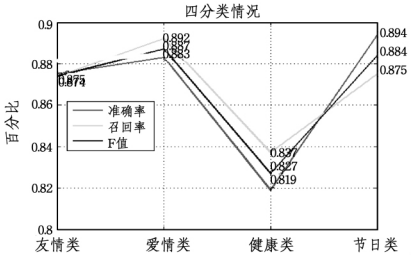

图 8　采用关联强度特征的半监督关联矩阵的分类情况

通过表 11 和图 8 可以得到，关联强度特征的半监督关联矩阵的综合准确率为 0.867，综合召回率为 0.869，综合 F 值为 0.868。

同时也可以看出，对于健康类短信，其准确率和召回率有所降低，这是由于在初始训练时健康类短信训练集较少，存在

分类模型不够准确的情况，而半监督学习扩大了这个错误。为了解决上述问题，系统再次采用了结合主动学习的方法，重新选取了 400 条由人工主动标注好的短信并通过算法 3 来重新修订关联矩阵 M，得到新的分类器。

然后再利用表 9 的测试集对其进行验证，实验结果如表 12 和图 9 所示。

表 12 采用关联强度特征的半监督关联矩阵结合主动学习

	友情类	爱情类	健康类	节日类	综合
准确率	0.902	0.921	0.892	0.878	0.898
召回率	0.910	0.897	0.863	0.902	0.893
F 值	0.905	0.908	0.877	0.889	0.894

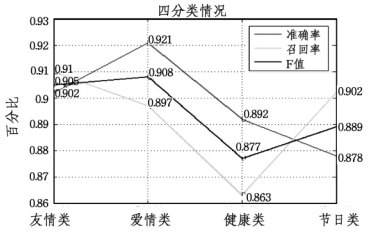

图 9 采用关联强度特征的半监督关联矩阵结合主动学习的分类情况

通过表 12 和图 9 可以得到，使用关联强度特征的半监督关联矩阵结合主动学习算法的综合准确率为 0.898，综合召回率为 0.893，综合 F 为 0.894。

3.4 实验结果分析

实验 1 是为了验证算法的可行性，并且与前人算法进行比较得出算法适用于短信自动分类；实验 2 的目的是验证算法的健壮性，检验该算法是否仅仅适用于类别比较少的情况。实验 2 证明，该算法在多分类情况下也有较好的分类效果。

综合实验 1 和实验 2，基于关联强度特征的半监督关联矩阵可以有效实现短信分类，再辅以半监督学习、半监督结合主动学习等措施，使得系统可以通过自我学习变得“更加聪明”。但本文研究依然存在着半监督学习对初始分类模型错误敏感度较高等问题，还需进一步解决。

实验 1 和实验 2 的训练集和测试集都是随机按 4∶1 来进行实验，但是考虑到其他比例的影响，本文也在实验中采取了多折交叉的方法对算法进行了验证，随机选取训练集和测试集的比例分别为 7∶3，6∶4，5∶5；同时也通过实验验证了关联强度和 $co(t_i,c_j)$，$Terms(c_j)$，Cf 以及 $D(c_j)$ 的关系等。但考虑到篇幅问题，本文没有列出以上各种情况。

结束语 本文提出了基于关联强度的关联矩阵概念，设计了全监督学习、半监督学习以及结合主动学习的各类算法。该系列算法可以有效实现系统的“自我学习”，对于短信中出现“新词”，也可以通过自我学习，校正分类模型，完成对“新词”的处理，同时也适应于其他短文分类的应用场合。

本文提出的分类算法已经应用于实际产品开发，在具体系统中还结合了短信的许多其他因素，如短信接收人、收件人、接收短信的时间等[31]，使得短信分类准确性也得到了较大的提高。

参 考 文 献

[1] 杨柳，殷钊，滕建斌，等. 改进贝叶斯分类的智能短信分类方法[J]. 计算机科学，2014，41(10)：31-35.

[2] 李继刚. 短信自动分类技术研究与应用[D]. 上海：东华大学，2012.

[3] 王文霞，王春红. 短信文本分类技术的研究[J]. 计算机技术与发展，2016，26(5)：145-148.

[4] 张永军，刘金岭. 基于特征词的垃圾短信分类器模型[J]. 计算机应用，2013，33(5)：1334-1337.

[5] 张永军，刘金岭. 一种改进的高效贝叶斯短信文本分类器[J]. 南京师范大学学报(工程技术版)，2014(3)：70-74.

[6] CHEN T，KAN M Y. Creating a Live，Public Short Message Service Corpus：The NUS SMS Corpus[J]. Language Resources & Evaluation，2011，47(2)：299-335.

[7] FORMAN G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research，2003，3(2)：1289-1305.

[8] 王文霞. 基于贝叶斯文本分类算法的垃圾短信过滤系统[J]. 山西大同大学学报(自然科学版)，2016(3)：17-19.

[9] 张杰，陈怀新. 基于归一化词频贝叶斯模型的文本分类方法[J]. 计算机工程与设计，2016，37(3)：799-802.

[10] 董红斌，滕旭阳，杨雪. 一种基于关联信息熵度量的特征选择方法[J]. 计算机研究与发展，2016，53(8)：1684-1695.

[11] 朱丽，陆建峰. 基于主动学习的微博聚类分析[J]. 数据采集与处理，2016，31(3)：599-605.

[12] 商宪丽，王学东. 微博话题识别中基于动态共词网络的文本特征提取方法[J]. 图书情报知识，2016(3)：80-88.

[13] 侯旭东. 基于内容的短消息智能分析系统研究[D]. 重庆：重庆理工大学，2010.

[14] 李建磊，王光辉，高宁，等. 结合全局信息描述子的局部特征匹配算法[J]. 测绘科学，2016，41(7)：33-36.

[15] 俸世洲. 基于自编码神经网络的文本表示应用研究[J]. 电子测试，2016(19)：91-92.

[16] 尚海，罗森林，韩磊，等. 基于句义成分的短文本表示方法研究[J]. 信息网络安全，2016(5)：64-70.

[17] 史玉珍，吕琼帅. 基于进化模糊规则的 Web 新闻文本挖掘与分类方法[J]. 湘潭大学自然科学学报，2016，38(2)：99-103.

[18] 马廷淮，金传鑫，侯荣涛，等. 一种基于术语频率和卡方统计的文本分类特征选择方法：CN104346459A[P]. 2015.

[19] 郭飞，张永锋. 一种新的中文文本分类特征提取的研究[J]. 数学的实践与认识，2016(12)：125-129.

[20] 孟佳娜. 迁移学习在文本分类中的应用研究[D]. 大连：大连理工大学，2011.

[21] 郭东峰，王东起. 机器学习中文本分类处理研究[J]. 内江科技，2016，37(9)：115-116.

[22] 黄旭. 基于机器学习的汉语短文本分类方法研究与实现[D]. 哈尔滨：黑龙江大学，2016.

[23] 路同强，石冰，闫中敏，等. 一种用于微博谣言检测的半监督学习算法[J]. 计算机应用研究，2016，33(3)：744-748.

[24] 朱晓光，聂培尧，林培光. 基于监督学习的微博情感分类方法[J]. 计算机应用与软件，2015(8)：238-242.

[25] 郭飞，张永锋. 一种新的中文文本分类特征提取的研究[J]. 数学的实践与认识，2016(12)：125-129.

[26] 江华丽. 中文分词算法研究与分析[J]. 物联网技术，2016(1)：87-89.

[27] 郭华平，董亚东，毛海涛，等. 一种基于逻辑判别式的稀有类分类方法[J]. 小型微型计算机系统，2016，37(1)：140-145.

[28] 吴红梅，牛耘. 基于特征加权的蛋白质交互识别[J]. 计算机技术与发展，2016(2)：114-117.

[29] http://www.aizhufu.cn/duanxinku/column/77_3/1.html.

[30] 李帅，陈笑蓉. 改进卡方统计量的 BPNN 短文本分类方法[J]. 贵州大学学报(自然科学版)，2015，32(6)：83-87.

[31] 李国栋，李卫. 基于文本分类技术的垃圾邮件识别系统[J]. 微电子学与计算机，2004，21(6)：145-146.