网络出版时间：2016-05-09 14:32:52
网络出版地址：http://www.cnki.net/kcms/detail/51.1196.TP.20160509.1432.102.html

基于词频的类别相关的特征权重算法

张　羚，陆余良，杨国正
(电子工程学院 网络系，合肥 230037)

摘　要：在文本分类领域中，目前关于特征权重的研究存在两方面不足：一方面，对于基于文档频率的特征权重算法，其中的文档频率常常忽略特征的词频信息；另一方面，对特征与类别的关系表达不够准确和充分。针对以上两点不足，提出一种新的基于词频的类别相关特征权重算法(全称 CDF-AICF)。该算法在度量特征权重时，考虑了特征在每个词频下的文档频率。同时，为了准确表达特征与类别的关系，提出了两个新的概念：类别相关文档频率 CDF 和平均逆类频率 AICF，分别用于表示特征对类别的表现力和区分力。最后，通过与其他 5 个特征权重度量方法相比较，在三个数据集上进行分类实验。结果显示，CDF-AICF 的分类性能优于其他 5 种度量方法。

关键词：文本分类；文本表示；特征权重；文档频率；逆类频率
中图分类号：TP391.1

Categories-related term weighting method based on term frequency

Zhang Ling, Lu Yuliang, Yang Guozheng
(Dept. of Network, Electronic Engineering Institute, Hefei 230037, China)

Abstract: In the field of automatic text classification, previous studies related to different term weighting had some deficiencies. On the one hand, for term weighting algorithm based on document frequency, term frequency is normally ignored in calculating document frequency. On the other hand, the expression of the relationship between the terms and the categories is not accurate and adequate. This paper developed a novel term weighting related categories based on term frequency (CDF-AICF). The algorithm took document frequencies for each term count of a term into account while measuring the term weight. In order to accurately express the relationship between terms and categories, the paper proposed two new concepts i.e., are document frequency related to category(CDF) and average inverse class frequency(AICF) respectively, used to reflect the expressive ability of term and the distinguishing ability of term. Finally, comparing with five related different term weighting approaches on three datasets, the performance of CDF-AICF is superior than the other five approaches.

Key Words: Text classification; text representation; term weighting; document frequency; inverse category frequency

电子媒体的发展与进步使发布和获取信息变得快速和便捷，人们可以通过互联网获取大量的数据信息，这些数据将近 80%以文本的形式存储[1]。如此庞大的数据通过人工进行整理是不可能的，需要有效的技术支持[2]，因而文本的自动分类技术成为重要的研究领域之一[3]。其中，文本表示是文本分类中的重要环节[4]。使用向量空间模型，文本内容表示为特征空间中的一个向量，每个特征词的权重由该词的重要程度决定。这种特征权重机制在文本分类中十分重要[5]，因此，本文主要研究文本特征的权重量化问题。

1　特征权重算法相关研究

1.1　TF-IDF

目前最常用的特征权重算法是 TFIDF。该算法由 Salton 提出[6]。它由两部分组成，第一部分是基于文档内容的词语频率(TF)，即词语在当前文档中出现的次数。第二部分是基于文档空间的文档频率(DF)，即在文档空间中出现过该词语的文档数。词语频率体现了特征对当前文档的表现力，词频越高，越能表示文档的内容，对文档的表现力越强，。词语的文档频率体现了特征对文档的区分力，在越多的文档中出现的特征，对文档的区分力越弱，特征的区分力与文档频率成反比，因此在计算时采用的是逆文档频率(IDF)。TFIDF 的经典计算公式为

$$W(t_i,d_j)=tf_{ij}*idf_i=tf_{ij}*\log\frac{D}{d(t_i)}$$

其中 tf_{ij} 为特征词 t_i 在文档 d_j 中词频，idf_i 为 t_i 的逆文档频率，D 为文档空间中的文档总数，$d(t_i)$ 为出现 t_i 的文档数，即 t_i 的文档频率。

1.2　存在不足

TF-IDF 算法表示了两方面的关系，一方面是特征词与当前文档的关系，另一方面是特征词与整个文档空间的关系。这两方面关系的权重量化可以将该文档与文档空间中的其它文档区分开。但在文本分类问题上，文档表示的最终目的不是将当前文档与其它文档区分开，而是将当前文档与其它类别的文档区分开。因此，对于文本分类问题，在文本表示中还应该考虑一种关系，特征与类别的关系。然而，这种关系并未在 TF-IDF 算法中体现，这也是该算法运用在文本分类问题中的不足之处。

1.3　其他研究

针对 TF-IDF 的不足，研究者提出了许多改进算法[7~11]，在

作者简介：张羚（1991-），女，江西乐平人，硕士研究生，主要研究方向为数据挖掘理论及其应用（371444672@qq.com）；陆余良（1964-），男，教授，博导，主要研究方向为数据挖掘理论及其应用；杨国正（1981-），男，讲师，博士，主要研究方向为数据挖掘理论及其应用.

特征权重量化过程中考虑了特征与类别的关系。其中，Man 等人[7]提出的相关频率(RF)算法考虑了特征的正类文档频率和负类文档频率，定义了相关频率(RF)对特征进行权重量化。Wang 等人[8]提出了基于逆类频率的 ICF-based 算法考虑了特征的类别频率(CF)。Ren 等人[9]对逆类频率因子进行改进，提出了基于类别密度频率的算法 TF-IDF-ICSF。以上算法在一定程度上考虑了特征与类别的关系，但仍存在一些局限，RF 算法将类别分为正类和负类，未考虑多个类别，因此，主要应用于二分类问题，然而我们面对的分类问题大部分为多分类问题。ICF-based 算法中加入了逆类频率因子，但由于类别数目的限制，很多特征词的逆类频率是相同的，无法准确比较不同特征词的区分能力。TF-IDF-ICSF 对逆类频率因子进行了改进，可是当训练集趋于平衡时，改进的逆类频率因子(ICSF)将退化成逆文档频率(IDF)。

2 基于词频的类别相关特征权重算法(CDF-AICF)

本章提出了一种基于词频的类别相关特征权重算法(CDF-AICF)。大多数对特征量化的算法中，在计算文档频率时通常忽略词语频率[12]，特征词的文档频率指出现过特征词的文档数，其中并不考虑特征在文档中的词频。为了挖掘特征与类别的关系，本文在特征权重度量中充分利用了词语的频率信息，把文档频率分成在不同类别下的文档频率，然后在此基础上，把每个类别中的文档频率再次分成在不同词频下的文档频率。为了进一步说明这个概念，考虑表 1 中的样例数据集 sample。表 1 显示了一个包含 12 个文本、3 个类别、4 个特征词的样例数据集，这是一个平衡数据集，每个类别有 4 个文本。表 2 显示了数据集中的四个特征词在不同词频下每个类别的文档频率。

表 1 样例数据集 sample

文档	类别	内容
文本 1	1	猫 猫 猫
文本 2	1	动物 猫
文本 3	1	猫 猫 鱼
文本 4	1	猫 狗 鱼 鱼
文本 5	2	狗 狗 狗
文本 6	2	动物 狗
文本 7	2	动物 狗 狗 猫
文本 8	2	狗 狗 猫 猫 鱼
文本 9	3	动物 猫 狗 鱼
文本 10	3	狗 狗 猫 猫
文本 11	3	猫 狗 猫 鱼
文本 12	3	动物 猫 鱼

表 2 样例数据集中的 4 个特征词在不同词频下的文档频率

词	猫			狗			鱼			动物		
词	类	类	类	类	类	类	类	类	类	类	类	类
频	1	2	3	1	2	3	1	2	3	1	2	3
1	2	1	2	1	1	2	1	1	3	1	2	2
2	1	1	2	0	2	1	1	0	0	0	0	0
3	1	0	0	0	1	0	0	0	0	0	0	0
	4	2	4	1	4	3	2	1	3	1	2	2

2.1 类别相关文档频率——CDF

基于特征在每个类别中的文档频率的不同，我们提出用特征在类别中的文档频率表示特征与类别的关系，具体地，表示特征对类别的表现力。其基本思想与 TF 类似，TF 用特征在当前文档中出现的次数表示特征对当前文档的表现力，文档中出现特征的次数越多，特征对文档越重要，应赋予越高的权重。类似地，提出一个新的概念——类别相关文档频率(CDF)，CDF 用特征在类别中出现的次数（文档频率）表示特征对类别的表现力，类别中出现特征的文档数越多，特征对类别越重要，应赋予越高的权重。但存在一个问题，特征与类别的关系并不像特征与文档的关系那样直接，我们首先需要获得与特征关联最大的类别。然后用该类别下的文档频率作为特征的 CDF 值。

为了获得与特征关联最大的类别，利用了特征的词频信息，因为特征在不同词频下各类别的文档频率分布不同，与之关联最大的类别也可能不同，以表 2 中的特征词“猫”为例，对于词频为 2 的“猫”，其在类别 1，2，3 的文档频率分别为 1，1，2。可知，包含词频为 2 的“猫”的文档最有可能属于类别 3，即与词频为 2 的“猫”关联最大的类别为类别 3。而对于词频为 3 的“猫”，其在类别 1，2，3 的文档频率分别为 1，0，0。可知，包含词频为 3 的“猫”的文档最有可能属于类别 1，即与词频为 3 的“猫”关联最大的类别为类别 1。对于同一个特征词“猫”，不同的词频，与之关联最大的类别并不相同。

因此，按以下步骤获得特征的 CDF 值，首先根据特征的词频信息，获得特征在该词频下文档频率在类别中的分布，把文档频率最高的类别作为特征在该词频下与之关联最大的类别。在获得与特征关联最大的类别后，将该类别下的文档频率作为特征的 CDF 值。在这个过程中，存在两种特殊情况：a)最高的文档频率可能有多个，即与特征关联最大的类别有多个，在这种情况下，从中随机选择一个类别作为与特征关联最大的类别;b)特征的词频数之前未出现过，如文本 d 中存在特征“猫”，其词频为 4。该词频数未在 sample 中出现，在这种情况下，选择与当前词频数距离最小的已知词频数（此例为 3）作为获得特征 CDF 值的词频依据。

以下用一个例子说明如何得到特征词的 CDF 值。假设一篇文档 d，其中包含特征词“猫”，且“猫”在 d 中的词频为 1，依据表 2 ，我们可以得到文档 d 中“猫”的 CDF 值。

根据“猫”在 d 中的词频 1，从表 2 中获得“猫”在词频 1 下文档频率在类别中的分布，即 2，1，2，该词频下，文档频率最高为 2，所对应的类别为 1 和 3，我们从 1 和 3 中随机选择一个类别作为与特征关联最大的类别，这里选择类别 1。然后，依据“猫”在各类别中的文档频率分布，即 4，2，4。将最大关联类别 1 所对应的文档频率 4 作为文档 d 中“猫”的 CDF 值。

算法 1 描述了获得 CDF 值的过程

算法 1 获得 CDF 值

输入：特征词 t_i

词频 f_j

$DF(t_i, f_j)=\{df_{ij1}, df_{ij2}, ..., df_{ijn}\}$，$df_{ijk}$ 表示类别 k 中包含词频为 f_j 的特征 t_i 的文档数，n 为类别总数

$DF(t_i)=\{df_{i1}, df_{i2}, ..., df_{in}\}$，$df_{ik}$ 表示类别 k 中包含特征 t_i 的文档数，n 为类别总数

输出：$CDF(t_i, f_j)$

```
max=0
for l=1 to n do
  if( df_ijl >max)
     max= df_ijl
end
m=0,C[],
for l=1 to n do
  if( df_ijl >max)
     C[m]=l
     m++
end
maxc=random(0,m)
CDF(t_i, f_j) = df_i,maxc
```

2.2　平均逆类频率——AICF

逆类频率 ICF 常用于表示特征与类别的关系[8, 9]，具体地，表示特征对类别的的区分力。其基本思想与 IDF 类似，IDF 认为大多数文档中都出现的特征，对文档的区分能力弱，特征对文档的区分能力与其文档频率成反比。类似地，ICF 认为大多数类别都出现的特征，对类别的区分能力弱，特征对类别的区分能力与其类频率成反比。极端情况下，如果某特征仅在一个类别中出现，则其对类别的区分能力最强，如果某特征在所有类别中都出现，则其对类别的区分能力最弱。与 IDF 的计算公式类似，ICF 的计算公式如下：

$$ICF(t_i)=\log(1+\frac{C}{c(t_i)})$$

其中，C 表示类别的总数，$c(t_i)$ 表示出现过特征 t_i 的类别总数，$c(t_i)/C$ 表示 t_i 的类别频率，而 $C/c(t_i)$ 表示 t_i 的逆类频率，$c(t_i)$ 越小，逆类频率越大，$ICF(t_i)$ 值越大。

ICF 存在一些不足，考虑以下三个特征，t_1、t_2 和 t_3，它们在三个类别中的文档频率的分布分别为 1、30、30 和 15、30、30 以及 30、30、30，三个特征的类别频率相同，且为最大，即在所有类别中都出现。因此 ICF 给 t_1、t_2、t_3 赋予最低的权值，且三个权值相等，ICF 认为 t_1、t_2、t_3 都对类别无区分能力，但很显然，t_1 和 t_2 对类别能进行一定程度上区分，且 t_1 的区分能力大于 t_2。以上分析可知，ICF 存在两点不足，第一，对于在所有类别中都出现的特征，ICF 赋予最低的权值，认为其无区分能力。第二，对于类别频率相同的特征，ICF 无法比较其对类别的区分能力。

针对 ICF 的不足，对 ICF 进行改进，在一定程度上弥补了 ICF 的两点不足。这里再次利用了特征的词频信息，改进后的 ICF 称为平均逆类频率(AICF)，平均逆类频率涉及以下几个概念：

a)全局类别频率（global class frequency）。指出现过特征词 t_i 的类别总数，用 GCF(t_i)表示。

b)局部类别频率（local class frequency）。指出现过词频为 f_i 的特征 t_i 的类别总数，用 LCF(t_i, f_i)表示。

c)平均类别频率（average class frequency）。指局部类别频率的平均数，其计算公式为：

$$ACF(t_i)=\frac{\sum_{j=1}^{n} LCF(t_i, f_j)}{n}$$，n 为特征出现过的词频的总数

对于 sample 中的特征词“猫”，全局类别频率 GCF(猫)=3。词频为 1 的“猫”的局部类别频率 LCF(猫，1)=3，词频为 2 的“猫”的局部类别频率 LCF(猫，2)=3，词频为 3 的“猫”的局部类别频率 LCF(猫，3)=1，特征词“猫”的平均类别频率为：ACF(猫)＝(LCF(猫，1)+LCF(猫，2)+LCF(猫，3))/3=(3+3+1)/3=2.33

本文提出的平均逆类频率 AICF 的概念考虑了特征在不同词频下的局部类别频率，通过局部类别频率 LCF 计算特征的平均类别频率 ACF，用 ACF 代替逆类频率 ICF 中的全局类别频率 GCF 得到平均逆类频率 AICF。平均类别频率能更好地估计特征在类别中的分布，用平均类别频率计算的平均逆类频率能更准确地判断特征对类别的区分能力，AICF 的计算公式如下:

$$AICF(t_i)=\log\left(1+\frac{C}{ACF(t_i)}\right)$$

表 3 显示了样例数据集 sample 中的特征词“猫”和“动物”的全局类别频率 GCF、局部类别频率 LCF、平均类别频率 ACF 以及逆类频率 ICF 和平均逆类频率 AICF。我们可以看到，“猫”和“动物”的 ICF 相等，ICF 无法区别“猫”和“动物”对类别的区分能力。但 AICF 弥补了这一点，表 3 显示 AICF(猫)>AICF(动物)，说明“猫”的类别区分能力大于“动物”，这与人们的直观感受相符合。

表 3 “猫”和“动物”的类别频率及逆类频率

	全局类别频率	局部类别频率			平均类别频率	ICF	AICF
		1	2	3			
猫	3	3	3	1	2.33	0.301	0.359
动物	3	3	/	/	3	0.301	0.301

2.3　特征权重算法——CDF-AICF

通过以上分析，一方面，类别相关的文档频率 CDF 体现了特征对类别的表现力，另一方面，平均逆类频率 AICF 体现了特征对类别的区分力。两方面结合能够综合反映特征与类别的关系。因此本文使用 CDF 和 AICF 两个因子对特征权重进行量化。对于 CDF，由于其值变化范围较大，在进行权重量化时，根据算法 1 求得 CDF 值后，对其取对数。AICF 则直接通过 2.2 节中的公式计算得到。因此，提出的特征权重算法 CDF-AICF 在计算词频为 f_j 的特征词 t_i 时，计算公式如下：

$$W_{CDF-AICF}(t_i, f_j)=\log\left(CDF(t_i, f_j)\right)*AICF(t_i)$$

对于文档 d_k 中词频为 f_j 的特征词 t_i，其余弦归一化表示为

$$W^{norm}_{CDF-AICF}(t_i, f_j, d_k)=\frac{\log\left(CDF(t_i, f_j)\right)*AICF(t_i)}{\sqrt{\sum_{t_i\in d_k}\left(\log\left(CDF(t_i, f_j)\right)*AICF(t_i)\right)^2}}$$

3　实验

在文本分类中存在许多特征权重算法，如相关频率算法[7]、基于逆类频率的算法[8]、基于概率的算法[13]、基于类别密度频率

的算法[9]等。为了验证本文算法 CDF-AICF 的有效性，对比了其他 5 种特征权重算法。表 5 显示了 5 种特征权重算法的名称及计算公式。公式中 A、B、C、D 的含义见表 4。

表 4 特征 ti 与类别 c_k 的列联表

	c_k	$\overline{c_k}$
t_i	A	C
$\overline{t_i}$	B	D

A 为 c_k 类中包含特征 t_i 的文档数；B 为 c_k 类中不包含特征 t_i 的文档数；C 为负类 $\overline{c_k}$ 中包含特征 t_i 的文档数；D 为负类 $\overline{c_k}$ 中不包含特征 t_i 的文档数

表 5 5 种特征权重度量方法

特征权重度量方法	简称	计算公式
TFIDF[6]	TF-IDF	$tf*\log\left(1+\frac{N_d}{d(t_i)}\right)$
相关频率算法[7]	RF	$tf*\log\left(2+\frac{A}{\max(C,1)}\right)$
基于逆类频率的算法[8]	icf-based	$tf*\log\left(2+\frac{A}{\max(C,1)}*\frac{N_c}{c(t_i)}\right)$
基于概率的算法[13]	ProbBa	$tf*\log\left(1+\frac{A}{B}*\frac{A}{C}\right)$
基于类别密度频率的算法[9]	TF-IDF-ICSF	$tf*\left(1+\log\frac{N_d}{d(t_i)}\right)*\left(1+\log\frac{N_c}{CS(t_i)}\right)$ $CS(t_i)=\sum_{c_k}\frac{n_{c_k}(t_i)}{N_{c_k}}$

* TF 表示词语频率，IDF 表示逆文档频率，ICF 表示逆类频率，ICSF 表示逆类密度频率

本节详细介绍了实验中所使用的数据集以及实验过程中的具体步骤，最后展示了实验的结果，并对其进行简要分析。

3.1 数据集

实验过程中使用了两个语料库，它们的规模和类别不同。第一个语料库是来自网易的文本分类数据，以下简称为网易 163，该语料库共收集文本 24000 篇，包含 6 个类别的文本，每类 4000 个文档，是一个平衡数据集。另一个语料库是中科院谭松波博士整理的 TanCorpV1.0，以下简称为 Tan，共收集文本 14150 篇，该语料库分为两个层次，第一层 12 个类别，第二层 60 个类别，本文分别在语料库的两个类别层次上构建数据集。因此本文实验的数据集共三个，分别是网易 163、Tan-12 和 Tan-60，其中 Tan-12 和 Tan-60 来自相同的语料库 TanCorpV1.0 的不同类别层次。这些数据集都可以从网上下载到它们的原始格式，每类我们都按 9：1 的比例将其分为训练集和测试集。表 6-7 显示了数据集的概况。

3.2 特征选择

在特征选择之前，首先使用几个预处理步骤对文档空间中的所有文档进行标准化。其中包括使用哈工大的 LTP 中的分词模块对数据进行分词，依据停用词表去除停用词，保留长度为 2~4 的词语，以及去除包含数字的词语（如 2013 年）。

表 6 数据集网易 163 和 Tan-12

数据集		类别					
		Auto	Culture	Economy	Medicine	Military	Sports
网易	T	3600	3600	3600	3600	3600	3600
	P	400	400	400	400	400	400
Tan-12		财经	地域	电脑	房产	教育	科技
	T	737	135	2648	841	727	936
	P	82	15	295	94	81	104
		汽车	人才	体育	卫生	艺术	娱乐
	T	531	547	2524	1265	491	1350
	p	59	61	281	141	55	150

* ‘网’代表网易数据集，P 代表训练集，T 代表测试集

表 7 数据集 Tan-60

类别								
两性	乒球	人才创业	人才履历	人才应试	人才猎取	人才管理	人才薪金	人物
301	100	35	35	35	35	370	36	57
34	12	4	4	4	4	42	4	7
企业	保健	出版	医药	古董艺术	地域城市	地域美食	地域风俗	城建
147	562	43	344	45	63	28	42	68
17	63	5	39	6	8	4	5	8
培训	天文科学	就业	心理	招生	文学艺术	校园	棋牌	水上
18	152	131	56	114	137	203	45	84
3	17	15	7	13	16	23	5	10
汽车快讯	汽车政策	汽车百科	汽车行驶	消费	生命科学	田径	电子商务	电影娱乐
232	34	106	158	81	413	75	623	450
26	4	12	18	10	46	9	70	50
电脑游戏	电脑病毒	电脑科技	电脑网络	电脑软件	留学	私宅	篮球	组屋
91	567	516	465	383	60	389	865	228
11	64	58	52	43	7	44	97	26
综艺娱乐	网球	美学艺术	考古科学	考试	自然科学	舞台艺术	装修	证券
450	117	75	164	155	206	166	154	192
50	14	9	19	18	23	19	18	22
财富	金融	音乐娱乐	音乐艺术	足球	羽球			
17	240	450	65	1185	49			
2	27	50	8	132	6			

由于本文研究的是用特征权重机制提高分类效果，因此，不运用像 IG、MI、卡方等特征选择技术[14~16]作为特征选择标准。我们用每类中词语的文档频率进行简单的本地特征选择[9]。这种方法统计词语在每类中的出现次数并对其降序排列。通过对数据

集中的每类设置门限，从而移除一部分特征。使用门限的方法是为了创建一个高维的向量空间（包含大量罕见词）和一个相对低维的向量空间（移除部分罕见词），以便评估本文方法在高维和低维两个空间的性能。

高维和低维的向量空间通过在两个数据集网易 163 和 Tan-12 的训练集中的每个类设置门限 ρ =10 和 ρ =20 产生。在进行本地特征选择之前，网易 163 和 Tan 的训练集的特征空间大小分别为 199073 和 72641。在高维向量空间中，门限 ρ =10 排除了训练集里每个类中出现少于 10 次的词语。相对地，门限 ρ =20 产生相对低维的向量空间，它排除了训练集里每个类中出现少于 20 次的词语。通过设置门限 ρ =10 和 ρ =20，网易 163 和 Tan-12 的特征空间中的特征数分别从 21388 下降到 12488 以及从 10462 下降到 6094。因此，通过设置不同的门限，我们消除了不同程度的罕见词，从而在高维和相对低维的向量空间中验证本文方法的有效性。

3.3 分类器

在机器学习领域中，许多分类器都成功应用于文本分类，如朴素贝叶斯（NB）、支持向量机（SVM）、KNN、决策树（DT）等。在所有著名的分类算法中，SVM 被认为是鲁棒性最好、分类最准确的算法之一[17]。因此选择 SVM 作为本文的分类器。实验中使用的是林智仁教授开发的 libSVM[18]工具包，所有的参数都使用默认参数。

3.4 性能评估

用于评价分类性能的标准有准确率、召回率和 F_1-measure[19]等。F1-measure 是准确率和召回率的调和平均数。准确率 $P(c_k)$、召回率 $R(c_k)$、$F_1(c_k)$ 的定义如下：

$$P(c_k)=\frac{TP(c_k)}{TP(c_k)+FP(c_k)}$$

$$R(c_k)=\frac{TP(c_k)}{TP(c_k)+FN(c_k)}$$

$$F_1(c_k)=\frac{2P(c_k)R(c_k)}{P(c_k)+R(c_k)}=\frac{2TP(c_k)}{2TP(c_k)+FP(c_k)+FN(c_k)}$$

$TP(c_k)$ 是类别 c_k 中正确判定为类别 c_k 的测试文档，$FP(c_k)$ 是其它类错误地判定为类别 c_k 的测试文档，$FN(c_k)$ 是类别 c_k 中错误地判定为其它类的测试文档。

为了评估平均性能，计算各标准的宏平均和微平均。类别空间中准确率、召回率和 F_1-measure 的宏平均 P^M、R^M、$F1^M$ 计算如下：

$$P^M=\frac{1}{N_c}\sum_{k=1}^{N_c}P(c_k)$$

$$R^M=\frac{1}{N_c}\sum_{k=1}^{N_c}P(c_k)$$

$$F_1^M=\frac{1}{N_c}\sum_{k=1}^{N_c}F_1(c_k)$$

类别空间中准确率、召回率和 F_1-measure 的微平均 P^μ、R^μ、F_1^μ 计算如下：

$$P^\mu=\frac{\sum_{k=1}^{N_c}TP(c_k)}{\sum_{k=1}^{N_c}(TP(c_k)+FP(c_k))}$$

$$R^\mu=\frac{\sum_{k=1}^{N_c}TP(c_k)}{\sum_{k=1}^{N_c}(TP(c_k)+FN(c_k))}$$

$$F_1^\mu=\frac{2P^\mu R^\mu}{P^\mu+R^\mu}$$

3.5 实验结果

在本文中设计了两个实验环境，分别是高维向量空间和相对低维的向量空间。在两个环境中，将本文方法与其他 5 种特征权重算法相比较，从而评估我们提出的方法的性能。表 8~11 显示了实验结果，黑体加下划线的数字表明每类的最好 F_1 值以及最好 F_1 宏平均值。

表 8 在网易 163 上的 F1 值（ρ =10）

类别	特征权重度量方法					
	TFIDF	RF	icf-based	ProbBa	TF-IDF-ICSF	CDF-AICF
Auto	0.9617	0.8775	0.8867	0.9208	**0.9656**	0.9590
Culture	0.8005	0.0148	0.0148	0.1176	0.704	**0.8045**
Economy	0.9267	0.8523	0.8589	0.7971	**0.9489**	0.8357
Medicine	**0.9322**	0.6386	0.6403	0.7939	0.8864	0.9164
Military	0.7800	0.6195	0.6253	0.6076	0.7334	**0.9443**
Sports	**0.9848**	0.7578	0.7684	0.9234	0.9848	0.9835
F1 宏	0.8977	0.6268	0.6324	0.6934	0.8705	**0.9072**

表 9 在网易 163 上的 F1 值（ρ =20）

类别	特征权重度量方法					
	TFIDF	RF	icf-based	ProbBa	TF-IDF-ICSF	CDF-AICF
Auto	0.9468	0.9146	0.9146	0.9538	0.9632	**0.9786**
Culture	0.8603	0.5754	0.5868	0.8748	0.8155	**0.8867**
Economy	0.8638	0.7826	0.7866	0.8046	0.8390	**0.9275**
Medicine	0.928	0.8741	0.8803	0.8990	0.9324	**0.9424**
Military	0.7450	0.5998	0.6076	0.8647	**0.9466**	0.8460
Sports	0.9585	0.8814	0.8861	0.9678	**0.9744**	0.9731
F1 宏	0.8837	0.7713	0.7770	0.8941	0.9118	**0.9257**

表 10 在 Tan-12 上的 F1 值（ρ =10）

类别	特征权重度量方法					
	TFIDF	RF	icf-based	ProbBa	TF-IDF-ICSF	CDF-AICF
财经	0.8636	0.7692	0.7692	0.7515	**0.8837**	0.7852
地域	0.64	0.9032	0.9032	0.9333	0.5217	**0.9655**
电脑	**0.9680**	0.9322	0.9322	0.9233	0.9617	0.9459
房产	0.9947	**1**	**1**	0.9893	0.9842	**1**
教育	0.8757	0.8901	0.8786	0.8690	0.8433	**0.9036**
科技	0.9082	0.9514	0.9514	**0.9556**	0.8780	0.9468
汽车	0.9830	0.9655	0.9655	0.9572	0.9649	0.9661
人才	0.6595	0.6736	0.6526	0.6021	0.6451	**0.6868**
体育	**0.9982**	0.9928	0.9928	0.9892	0.9982	0.9946
卫生	0.9547	0.9611	0.9611	0.9510	0.9342	**0.9754**
艺术	0.7636	0.9107	0.9107	0.9541	0.7169	**0.9549**
娱乐	0.9240	0.9662	0.9662	0.9832	0.9009	**0.9932**
F1 宏	0.8778	0.9097	0.9069	0.9049	0.8527	**0.9265**

表 11 在 Tan-12 上的 F1 值（ρ=20）

类别	特征权重度量方法					
	TFIDF	RF	icf-based	ProbBa	TF-IDF-ICSF	CDF-AICF
财经	**0.8685**	0.7692	0.7692	0.7515	0.8639	0.7852
地域	0.64	0.9032	0.9032	**0.9333**	0.5833	0.9285
电脑	**0.9680**	0.9337	0.9337	0.9248	0.9569	0.9459
房产	0.9842	**1**	**1**	0.9893	0.9842	**1**
教育	0.8757	0.8901	0.8901	0.8690	0.8452	**0.9036**
科技	0.9073	0.9463	0.9463	**0.9556**	0.8932	0.9468
汽车	**0.9915**	0.9655	0.9655	0.9572	0.9739	0.9661
人才	**0.6875**	**0.6875**	**0.6875**	0.6170	0.6521	0.6734
体育	0.9964	0.9928	0.9928	0.9892	**0.9982**	0.9964
卫生	0.9510	0.9577	0.9577	0.9580	0.9415	**0.9754**
艺术	0.7850	0.9107	0.9107	**0.9541**	0.7572	0.9464
娱乐	0.9276	0.9662	0.9662	0.9832	0.9225	**0.9932**
F1 宏	0.8819	0.9102	0.9102	0.9069	0.8643	**0.9217**

3.5.1　高维空间的结果

高维空间通过在数据集网易 163 和 Tan-12 设置门限 ρ=10 产生。表 8 显示了在网易 163 中每类的 F_1 值及 F_1 的宏平均值。表中的数据说明，与其他的特征权重算法相比较，CDF-AICF 的 F_1 宏平均值最高，CDF-AICF 的总体分类的性能优于其它方法。

相比之下，表 9 显示了在数据集 Tan-12 中的分类结果，与其它特征权重方法相比，CDF-AICF 在 12 个类别中的 7 个达到最高的 F_1 值，且 F_1 的宏平均值最高，结果显示了 CDF-AICF 方法的良好性能。

高维空间的结果显示，CDF-AICF 特征权重算法优于其它方法。这一部分是因为 CDF-AICF 方法中的 AICF 因子赋予罕见词积极的区分能力，而高维的向量空间中包含大量的罕见词。因此，在高维向量空间中，CDF-AICF 特征权重算法表现出它的优势。

3.5.2　低维空间的结果

低维空间通过在数据集网易 163 和 Tan-12 设置门限 ρ=20 产生。表 10 和表 11 显示了低维空间的结果。其中，表 10 显示了在网易 163 的结果，CDF-AICF 在 6 个类别中的 4 个达到最高的 F_1 值，且 F_1 的宏平均值最高。

相比较，表 11 中的数据显示了在数据集 Tan-12 的结果，与其它的特征权重算法相比，CDF-AICF 的 F_1 宏平均值最高，CDF-AICF 的总体分类的性能优于其它方法。

以上结果显示，在低维空间中，CDF-AICF 特征权重算法也优于其它方法。

3.5.3　数据集 Tan-60 的实验结果

在这个数据集中，各类别的文本分布极不均衡，表 7 显示了数据集 Tan-60 在各个类别上的文本数，其中大多数类别的文本数很少，一些类别的训练文本甚至只包含十几个文本。因此，对于 Tan-60 数据集，不是对训练集的每个类设置门限，而是对训练集整体设置门限 ρ=3，排除在整个训练集中出现次数少于 3 的特征，排除极其罕见词，特征空间的特征总数从 72641 下降为 21408。图 1 显示了各个权重算法在数据集 Tan-60 上的 F_1 微平均值。结果显示，CDF-AICF 的 F_1 微平均值最高，为 79.43，ICF-BASED 和 RF 名列第二和第三，分别为 78.95 和 76.77。

3.5.4　总体性能

图 1 用 F_1 微平均值分别在三个数据集上比较了 5 个不同的特征权重方法的分类性能。从图中的结果可知，在三个数据集上，CDF-AICF 的表现都优于其它特征权重方法。

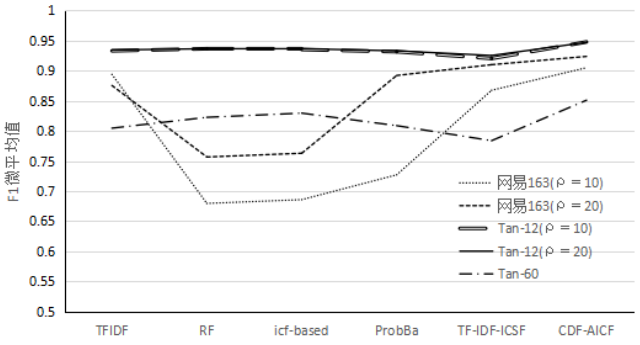

图 1 三个数据集上的 F1 微平均值

以上实验结果显示，提出的 CDF-AICF 特征权重算法在不同的空间维度上性能稳定，在不同的数据集上表现也优于其它方法。除此之外，CDF-AICF 不仅在体现整体分类性能的 F_1 微平均和 F_1 宏平均上表现出色，而且在各个类别中的 F_1 值也位居前列。这是因为本文考虑到在分类过程中，各文本是以类别为单位与其他文本区分开，因此在特征权重量化时充分表现特征与类别的关系。实验结果验证了该方法的有效性。此外，将词频信息融入其中，一方面充分地利用了特征的信息，另一方面也更准确地表现了特征与类别的关系。但不足之处是：在加入词频信息的同时，增加了算法的复杂度。与其它方法相比，本文方法需统计每个词频下的文档频率，因此，计算特征权重时耗时更长。

综上所述，本文算法的优势在于性能稳定，分类效果理想。

4 结束语

本文首先充分利用特征的词频信息，考虑了特征词在每个词频下的文档频率。然后，在此基础上，提出了两个与类别相关的概念 CDF 和 AICF。最后，提出一种新的特征权重度量算法——基于词频的类别相关特征权重度量算法(CDF-AICF)。我们通过对比其它 5 种特征权重度量算法，使用 SVM 分类器，在三个数据集上进行实验，实验结果表明，提出的 CDF-AICF 算法的分类效果优于其它 5 种特征权重度量方法。

从目前工作上看，CDF-AICF 算法在单标签的分类任务中表现不错，在下一步的研究中，将应用该算法到多标签的分类任务中，研究其是否能处理多标签的分类及其性能如何。同时，还可以结合语义信息丰富特征对文本的表达，进一步提高文本分类的性能。

参考文献

[1] Harish B S, Guru D S, Manjunath S. Representation and Classification of Text Documents: A Brief Review[J]. **International Journal of Computer Applications**, 2010, **8**(2):110-119.

[2] Tang P, Chow T W S. Mining language variation using word using and collocation characteristics[J]. **Expert Systems with Applications**, 2014, **41**(17):7805–7819.

[3] Baharum B, Lee L H, Khairullah K. A review of machine learning algorithms for text-documents classification[J]. **Journal of Advances in Information Technology,** 2010, 1(1):4-20.

[4] Guo Y, Shao Z, Hua N. Automatic text categorization based on content analysis with cognitive situation models[J]. **Information Sciences**, 2010, **180**(5):613-630.

[5] Li W, Miao D, Wang W. Two-level hierarchical combination method for text classification[J]. **Expert Systems with Applications**, 2011, **38**(3):2030-2039.

[6] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. **Information Processing & Management**, 1988, **24**(5):513-523.

[7] Man L, Chew Lim T, Jian S, et al. Supervised and traditional term weighting methods for automatic text categorization.[J]. **IEEE Trans on Pattern Analysis & Machine Intelligence**, 2009, **31**(4):721-735.

[8] Wang D, Zhang H. Inverse-category-frequency based supervised term weighting schemes for text categorization[J]. **Journal of Information Science & Engineering**, 2013, **29**(2):209-225.

[9] Ren F, Sohrab M G. Class-indexing-based term weighting for automatic text classification[J]. **Information Sciences**, 2013, **236**(1):109-125.

[10] Zhang H, Wang D, Wu W, et al. Term frequency–function of document frequency: a new term weighting scheme for enterprise information retrieval[J]. **Enterprise Information Systems**, 2012, **6**(4):433-444.

[11] Rehman A, Javed K, Babri H A, et al. Relative discrimination criterion：a novel feature ranking method for text data[J]. **Expert Systems with Applications**, 2015, **42(**7):3670-3681.

[12] Baccianella S, Esuli A, Sebastiani F. Using micro-documents for feature selection: The case of ordinal text classification[J]. **Expert Systems with Applications**, 2013, **40**(11):4687-4696.

[13] Liu Y, Han T L, Sun A. Imbalanced text classification: A term weighting approach[J]. **Expert Systems with Applications**, 2009, **36**(1):690–701.

[14] Covões T F, Hruschka E R. Towards improving cluster-based feature selection with a simplified silhouette filter.[J]. **Informationences,** 2011, **181**(18):3766-3782.

[15] Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines[J]. **Information Sciences**, 2011, 181(1):115-128.

[16] Ogura H, Amano H, Kondo M. Comparison of metrics for feature selection in imbalanced text classification[J]. **Expert Systems with Applications,** 2011, **38**(5):4978-4989.

[17] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. **Knowledge & Information Systems**, 2007, **14**(1):1-37.

[18] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. **ACM Trans on Intelligent Systems & Technology**, 2011, **2**(3):389-396.

[19] Yang Y. An evaluation of statistical approaches to text categorization[J]. **Information Retrieval**, 1999, **1**(1-2):69-90.