基于类别分析及有效特征提取的文本分类方法

李云　盛艳　栾鸾
(扬州大学 信息工程学院，江苏扬州 225009)
E-mail: shengyan1985@gmail.com

摘要: 文本分类是指在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。为了提取更少的特征来尽可能的表达出文本蕴含的信息，本文首先通过分析各个特征自身的统计特性，根据Zipf定律进行全局特征提取，不提取文本特征空间中普遍存在的特征和噪声特征；其次，在对特征的类别信息进行统计分析后，计算出每个特征词的类别贡献程度，进行类内局部特征提取；在选取有效的特征之后，利用类别频率对传统的TF-IDF公式进行改进，提出新的特征权重计算公式TF-IDF-CF；接着，基于前两步骤获得的结果进行文本分类；最后在数据集Newsgroup上进行测试。实验证明本文所采用的特征提取方法是合理的，改进的权重计算公式也能够得到较高的分类准确度。
关键词: 特征提取；Zipf 定律；类别频率；特征权重

Text Feature Extraction Method and Improved TF-IDF Formula based on Category Analysis

Li Yun　Sheng Yan　Luan Luan
(Institute of Information Engineering，Yangzhou University，Jiansu Yangzhou 225009)
E-mail: shengyan1985@gmail.com

Abstract: Text classification refers to determine the class of an unknown text according to its content in the given classification system. In order to extract fewer features to express the information in the text as much as possible, the paper analysis the various features' statistical properties and to extract the features in the whole according to Zipf's law firstly; Secondly, based on the statistical analysis of the features' classified information, we compute the contribute of a feature and extract the efficient feature; After that, we improve the traditional TF-IDF formula using category frequencies and put forward a new formula for calculating the feature weight named by TF-IDF-CF; Next, Text classification method is advanced on the base of the first two steps; Finally test the algorithm in the Newsgroup data set. This experiment result proved that feature extraction methods advanced in the paper are reasonable and the formula for calculating the feature weight has higher classification accuracy.
Key word: Feature Extraction; Zipf Law; Category Frequencies; Feature Weight

1 引 言

现代网络不断发展的同时，资源信息尤其是文本信息，也日益膨胀，人们需要投入更多的时间对信息进行组织和管理。对这些庞大的信息进行人工分类是相当耗费精力的，所以利用计算机进行信息的自动分类技术已成为数据挖掘领域中一个的重要研究方向，并且具有很高的商业价值。简单来说，文本分类[1]的任务是：在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。现有的大多数文本分类系统都使用了向量空间模型对文本进行表示，即是把一个文本看成是特征词序列，并计算这些特征词的权重，将文本表示成这些权重的向量形式，然后再对这个向量空间进行处理。

由于一个文本包含的特征词往往非常多，所得到的向量空间的维数非常大，这么高维的向量空间使得文本分类的时间、空间消耗急剧上升。因此，一般在进行文本分类之前，需要对文本进行特征提取和特征降维，那么，如何做到在没有失去可用信息的情况下提取出更少

的特征，尽可能得让特征维数更小呢？很多文献[2~9]提出了不同的方法来特征提取和特征降维。文献[2~4]对传统的 TF-IDF 计算方法进行改进，其中文献[4]提出一种基于词频差异的特征提取方法，以提高特征提取质量和文本分类的准确度。文献[5]对 CHI 公式进行改进来更好的表示出特征对类别的贡献程度和相关程度。文献[6]主要集成多种特征提取方法对关系密切的特征进行合并以达到降维的目的。文献[7~9]在对类别信息进行分析之后，根据特征的类别区分能力进行特征选择，选择出对文本分类中具有较高类别区分能力的特征。从中可以看到，为了能够得到更好的特征向量空间，可以通过以下 2 个方法：1）根据各个特征自身的统计特性，过滤掉普遍的特征，去除噪声，以提取出更有意义的特征；2）可以对文本类别信息进行分析，计算出每个特征词的类别贡献程度，从而能够在文本自动分类中得到更准确的效果。然而，文献[2~9]中大多是从单一角度出发，即或是对传统公式进行改进，或仅仅利用特征类别贡献度，又或是从词频特性中进行特征提取而对后续的分类过程没有处理，而整个过程需要经过前续的特征提取和后续的分类两个步骤。如果特征提取过程中提取的特征非常符合后续分类的需要，那么总的效果就会很好，反之效果则相对较差。同样，如果仅对后者改进而前者没有，那么总的结果也不会很理想，所以应该将两者结合起来统一考虑。

基于这个思想，本文首先根据 Zipf[10~13]定律进行全局特征提取，不提取整个文本特征空间中普遍存在和特别稀少的噪声特征；其次，对特征的类别信息进行统计分析后，进行类内局部特征提取；在选取有效的特征之后，利用类别频率对传统的 TF-IDF 公式进行改进，最终得到一个全面统一的分类方法。本文最后在数据集 NewsGroup 上进行测试，证明采用的特征提取方法是合理的，分类方法也能够得到较高的分类准确性。

2 背景知识

2.1 齐普夫定律（Zipf's Law）

1932 年，哈佛大学的语言学专家 Zipf[10]在研究英文单词出现的频率时，发现如果把单词出现的频率按由大到小的顺序排列，则每个单词出现的频率与它的名次的常数次幂存在简单的反比关系：

$P(r)=\frac{C}{r^{\alpha}}$，这种分布就称为 Zipf 定律（Zipf's law）。其中，C 为一个正的常数，通常为 1，α 也为一个正的常数，称为 Zipf 指数，大小仅取决于具体的分布，与其他参数无关，在英语中约为 1。这一规律表明在语言中经常被使用的词汇只占词汇总量的很少一部分，而绝大部分词汇则很少被使用。它的频率-名次曲线[13]如图 1 所示。如果对词频及其序号做对数-对数（log-log）曲线[13]，将会出现一条斜率约为 $-\alpha$ 的直线。

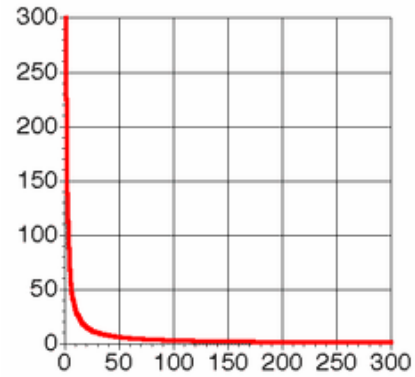

图 1 zipf 定律频率-名次曲线示意图

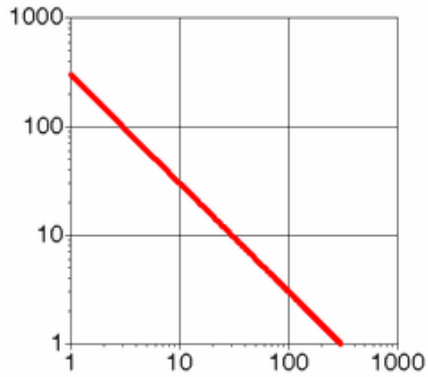

图 2 zipf 定律对数-对数曲线示意图

Zipf 定律表明在英语单词中，只有极少数的词被经常使用，而绝大多数词很少被使用。

实际上，许多人类语言都有符合这个定律。从 Zipf 定律分布曲线图中，可以看到，频率高的词的个数不多，大多数的词很少被使用。本文正是利用文本词集中的这种特性来进行特征提取的。

2.2 传统特征提取方法

目前有很多种特征提取算法[5]，如文档频率（DF），信息增益（Information Gain，IG），互信息（Mutual Information，MI），χ^2 统计法（CHI）。下面分别简单介绍下各个方法[5]：

1) 文档频率（Document Frequency，DF）。 文档频率是指在训练文本集中某一特征词出现文本数。采用 DF 作为特征抽取基于如下基本假设：DF 值低于某个阈值的词条是低频词，它们不含或含有较少的类别信息。将这样的词条从原始特征空间中删除，不但能够降低特征空间的维数，,而且还有可能提高分类的精度。

DF 的优点在于计算量很小，而在实际运用中却有很好的效果。缺点是稀有词可能在某一类文本中并不稀有，也可能包含着重要的判断信息，简单舍弃可能影响分类器的精度。

2) 信息增益（Information Gain，IG）。信息增益是一种基于熵的评估方法，涉及较多的数学理论和复杂的熵理论公式，定义为某特征项为整个分类所能提供的信息量，不考虑任何特征的熵与考虑该特征后的熵的差值。它根据训练数据，计算出各个特征项的信息增益，删除信息增益很小的项，其余的按照信息增益从大到小排序。信息增益计算公式如下：

$$IG(t) = P(t)\sum_{i=1}^{M} P(C_i \mid t)\log\frac{P(C_i \mid t)}{P(C_i)} + P(\overline{t})\sum_{i=1}^{M} P(C_i \mid \overline{t})\log\frac{P(C_i \mid \overline{t})}{P(C_i)}$$

其中 t 表示特征项, P(t)表示训练集中包含特征项 t 的文本的概率，$P(C_i)$表示类别 C_i 在训练集中出现的概率，$P(C_i|t)$表示文本包含特征项 t 时属于 C_i 类的条件概率。$P(\overline{t})$ 表示训练集中不包含特征项 t 的文本的概率，$P(C_i \mid \overline{t})$ 表示文本不包含特征项 t 时属于 C_i 类的条件概率。显然，某个特征项的信息增益值越大，类别区分能力就越强。

从信息论角度出发，信息增益方法的本质是用各个特征值来划分训练样本空间，根据所获信息增益的多少来选择相应的特征。不足之处在于，它考虑了词未出现的情况。虽然某个词不出现可能对判断文本类别也有贡献，但实验证明这种贡献往往小于考虑词不出现情况所带来的干扰，因为一篇文本仅能包含特征空间中的很少一部分特征，此时信息增益大的特征主要是信息增益公式中后一部分（代表单词不出现情况）大，而非前一部分（代表单词出现情况）大，信息增益的效果就会大大降低。

3) 互信息（Mutual Information ，MI）。互信息是一种广泛用于建立特征项关联统计模型的标准，它体现了特征项与类别的相关程度。对于特征项 t 和某一类别 C_i（i = 1，2，…，m），在 C_i 中出现的概率高，而在其他类别中出现的概率低的特征项 t 将获得较高的互信息，也就有可能被选取为类别 C_i 的特征。互信息的估计值定义为：

$$\text{Mutual Info Txt}(t,\ C) = \log\frac{A\times B}{(A+C)(A+B)}$$

其中，A 表示特征 t 与类 C_i 同时出现的次数，B 表示特征 t 不在类 C_i 中出现的次数，C 表示类 C_i 中没有出现特征 t 的文本数。

互信息的一个好处是，既能反映类与特征之间的关系（相对于 DF），计算又不至于太复杂（相对于 CHI 和 IG），缺点是没有考虑 C 特征发生的频率，这造成了互信息评估函数经常倾向于选择稀有单词，从而淘汰了很多高频的有用词条。

4) χ^2 统计法（CHI）。χ^2 统计法用于度量特征和类别之间独立性的缺乏程度，它同时

考虑了特征存在与不存在的情况。χ^2 越大，独立性越小，相关性越大；反之独立性越大，相关性越小。

χ^2 统计量表示为：

$$\chi^2(t,\ C)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

$$\chi^2_{AVG}(t)=\sum_{i=1}^{m}P(c_i)\chi^2(t,c_i)\text{(对所有类别求和)}$$

$$\chi^2_{Max}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\}\text{(取词条对所有类别的CHI最大值)}$$

其中，A 为 t 和 c 同时出现的次数，B 为 t 出现而 c 没有出现的次数，C 为 c 出现而 t 没有出现的次数，D 为 t 和 c 都没有出现的次数，N 为训练集中所有实例文本数。

χ^2 统计量可以用来度量类 c 和特征 t 的关联性，使得它在特征削减中十分有用。对每一对 c 和 t 都计算 χ^2 的值，并按照降序排列，去除排在后面的特征。χ^2 统计量与互信息的差别在于它是归一化的统计量，但是它对低频特征项的区分效果也不好。

2.3 特征权重计算

特征权重的计算方法对应各种特征选择方法也有很多种不同的算法，常用的有布尔值法、词频法、TF-IDF 法。基于向量空间模型中的特征权重计算方法一般采用 TF-IDF 公式。

传统的 TF-IDF 公式，即如下所示：

$\omega(t,d)=\dfrac{TF(t,d)\times\log(IDF(t)+\delta)}{\sqrt{\sum_{t\in d}(TF(t,d)\times\log(IDF(t)+\delta))^2}}$，其中，TF（t，d）为特征 t 在文本 d 中的频率，IDF（t）为逆文档频率，文档频率 DF（t）的倒数，即 $IDF(t)=\dfrac{1}{DF(t)}$，δ 为调节系数，一般为 0.01。用 TF-IDF 算法来计算特征的权重值是表示当一个特征词在这篇文档中出现的频率越高，同时在其他文档中出现的次数越少，则表明该特征对于表示这篇文档那个的区分能力越强，所以其权重就应该越大。也可以从这个公式中看到特征 t 的权重只根据词频和文档频率，没有考虑到它的类别频率，而往往类别频率在文本分类中也起着一定作用，所以本文提出一个改进的特征权重计算方法。

3 特征提取策略

由于特征提取的效果好坏直接制约着后续的文本分类效果，而现有的特征提取仅从单一层面出发，没有从全局角度进行提取。所以本文提出了一种新颖的特征提取方法，他分为两个步骤，分别是全局特征提取和局部特征提取，前者针对整个训练文本集（Training Corpus，简记为 TC），利用 Zipf 定律去除普遍存在的和特别稀少的噪声特征，实现在全局范围内的特征提取；而后者是在对每个特征词的类别信息分析之后，根据得到的类别频率进行类内的局部特征提取。在这个过程中，除了计算每个特征的词频 TF，文档频率 DF 外，还需计算特征的全局频率（Overall Frequency，OF）用于全局特征提取和特征的类别频率（Category

Frequency，CF）用于局部特征提取。

3．1 全局特征提取

由于人类自然语言中普遍存在 Zipf 定律，即只有极少数的词被经常使用，而绝大多数词很少被使用。对于普通文本集中，同样呈现这种规律，那么在提取特征时，不提取频率很高的少数特征，同时对于那些频率极低的特征，很可能是噪声的特征，将其从原始特征空间中删除。

定义 1：在整个训练文本集 TC 中，特征词 t 在 TC 中出现的频率称为特征词 t 的全局频率 OF，定义为：$OF(t)=\frac{t\text{在TC中出现的次数}}{\text{TC中所有词数}}$。

在计算出所有特征的全局频率之后，按照全局频率从大到小进行排序得到一张频率-位置分布图，大多数情况下它是符合 zipf 定律的。因此我们可以通过设定两个阈值：全局低频特征百分比 z-low，全局高频特征百分比 z-high，它们分别用于删除在频率-位置分布图中所有特征个数×z-low 个较低频率特征和所有特征个数×z-high 个较高频率特征，即只保留整体频率在频率-位置分布图中位于两个阈值之内的特征。另外，经过计算每个特征的文档频率 DF 之后，如果某个特征的 DF 值低于一定阈值，也在这个步骤中将其从特征空间中删掉。这个过程可以由算法 1 来描述：

算法 1: Overall_Selected(TC，z-low，z-high，df-threshold)

输入: 训练词集 TC，全局低频特征百分比 z-low，全局高频特征百分比 z-high，DF 阈值 df-threshold

输出: 有效特征集 TSet

描述:

Step 1：根据 TC，获得所有特征集 TSet 并计算每个特征 t 的全局频率

Step 2：对 TSet 根据全局频率大小，从高到低排序

high_OF = TSet 中从位置 1 到位置$\lfloor |TSet|\times z-high \rfloor$的特征集 # 待删除的全局高频特征集

low_OF = TSet 中从位置$\lfloor |TSet|\times z-low \rfloor$到位置$|TSet|$的特征集 # 待删除的全局低频特征集

特征集 TSet 中删掉 high_OF 和 low_OF 中的特征

Step 3：# 在文本向量空间中删除全局高频/低频特征, 并计算每个特征的 DF

```
for TC 中的每个文本 file 依次进行：
    for 文本 file 中的每个特征 term 依次进行：
        if term in high_OF or term in low_OF：
            将 term 从文本 file 中删除
    for 特征集 TSet 每个特征 t：            # 计算 DF
        if 这个特征 t 在该文本 file 中：
            对应特征 t 的频数增 1
```

Step 4：# 去除 DF 低的 term

根据 TSet 中每个特征 t 的频数转换计算 DF

在文本空间中删除 DF 小于 df-threshold 的特征并返回最终的特征集 TSet

这个算法是从全局出发，在整个文本特征空间中过滤掉出现次数非常频繁的，或出现次数极少（很有可能是噪声）的特征，能够减少原始文本特征空间中的无效特征个数。

3．2 局部特征提取

每个特征在文本特征空间中具有特定的类别区分能力，这些类别信息对文本特征提取也是非常有意义的。因此，可以统计每个特征词在各类别中的出现频率，这里称为类别频率，其定义如下：

定义 2：在整个训练文本集 TC 中，其类别集合$C=(c_1,c_2,...c_m),m$为类别数，每个文本已指明类别信息$c_i \in C$，对于特征词 t 在文本对应的类别$c_i \in C$中出现的频率称为特征词 t 的类别频率 CF，定义为：$CF(t,c_i)=\frac{t在类c_i中的出现次数}{类c_i中的所有词数}$。

对于特征 t，如果它在较少的类别中出现且在较少的类别中出现的频率较高，表示这个特征在这些类别中是特定含义的，它相对于其他特征来说具有较大的区分能力，所以像这样的特征应该尽可能保留下来;但如果它在大多数的类别中都出现并且它在各个类别中的类别频率都非常接近，那么说明特征 t 对类别区分能力不大，所以可以把该特征从特征空间中删除。最终保留下来的是类别频率差异较大的特征，在此提出算法 2 进行特征空间的局部提取。

算法 2: Category_Selected(TC，cf-rate，cf-threshold)

输入: 训练词集 TC，类别出现率 cf-rate，词频差异阈值 cf-threshold

输出: 有效特征集

描述:

```
    Step 1：# 计算特征 t 的类别频率 CF
            for 对于 TC 的每个文本 file 依次进行:
                获得该文本 file 所在的类别 c
                for 对于文本 file 中的每个特征 t 依次进行:
                    CF(t,c)增 1
            根据每个特征 t 在各类别中的频数转换计算得到特征的 CF
    Step 2：CF_min = 总的类别数*cf-rate          # 即是类间频率阈值
            for CF 中的每个特征 t 依次进行:
                if |特征 t 的类别数| >= CF_min and _IsInThreshold(特征 t 的类别频率,cf-threshold):
# 如果特征 t 在大多数类别中出现并它的类别频率值相差不大，就把 t 删除
                    删除特征 t
    Step 3：返回最终的特征集
```

_IsInThreshold(cf，cf-threshold)

输入: 一系列类别频率值 cf，词频差异阈值 cf-threshold

输出: 数值波动是否在 cf-threshold 内, 如果是返回 True, 否则 False

描述:

```
    Step 1: # 寻找出类别频率中的最大值和最小值
            cf_min 为类别频率 cf 中的最小值，初始化为 cf 中的第一个数值
            cf_max 为类别频率 cf 中的最大值，初始化为 cf 中的第一个数值
            for 每个剩余 cf 中的 ocf 依次进行:
               if otc < tc_min:
                   tc_min = otc
                   continue
```

```
            if otc > tc_max:
                tc_max = otc
Step 2：#判断波动是否在给定词频差异阈值内，如果满足则返回 True，否则返回 False
        if tc_max-tc_min < CF_THRESHOLD:
            return True
        return False
```

文本特征空间在经过算法 2 的处理之后，得到的是具有类别区分能力较大的特征集，而且这个特征集中的特征个数要比原始的特征集中的要小的多。

这样，经过全局特征提取和局部特征提取之后，能够在尽可能保留原始文本特征信息的情况下，最大化的降低特征维数，从而提高文本自动分类效率。

4 基于类别分析和有效特征的文本分类算法

4.1 基于类别频率的特征权值计算公式

依据 2.3 的传统 TF-IDF 公式，编程很容易实现且计算方便，效率也较高，所以被广泛应用于很多实际分类和检索系统中，这是 TF-IDF 的最大优点。但是从中也可以看到，它的特征权重只根据词频和文档频率，没有考虑自身的类别频率，而往往类别频率在文本分类中也起着一定作用，所以本文提出一个改进的特征权重计算公式，记为 TF-IDF-CF：

$$\omega(t,d)=\frac{TF(t,d)\times\log(IDF(t)+\delta)\times\max\limits_{c\in C}CF(t,c)}{\sqrt{\sum\limits_{t\in d}(TF(t,d)\times\log(IDF(t)+\delta)\times\max\limits_{c\in C}CF(t,c))^2}}$$，其中，TF（t，d）为特征 t 在文本 d 中的频率，IDF（t）为逆文档频率，δ 为调节系数，一般为 0.01。CF（t，c）为特征 t 在类 c 中的类别频率， 其计算方法和算法 2 中相同，由于特征 t 在类别集合 C 中对应多个类别可能有多个类别频率，为了更好的反映出特征 t 的类别贡献度，如果特征 t 的某个类别频率越大，就表示其类别贡献度越大，反之越小，所以应该在所有的类别频率中取最大值来加重特征 t 在文档 d 中的权重。从全局看，公式中加入类别频率 CF 之后，可以看到如果特征 t 的类别频率越高，相应的特征权重也就越高，表示特征 t 越有意义。反之，如果特征 t 的类别频率越低，所得到的特征权重就越小。

4.2 改进 kNN 分类算法

使用改进的权值计算公式 TF-IDF-CF 对算法 2 中提取的有效特征进行权重计算之后，提出一种类似于传统分类算法 kNN[14]的算法 3 来进行文本分类。主要思想是，利用算法 2，对整个训练文本集提取所有有效的特征之后，根据 4.1 中提出的特征权值计算公式计算特征权重，从而构成整个文本向量空间。接着，就对未知文档以同样的方式提取有效特征并用 TF-IDF-CF 计算特征权重得到一个文本向量，在整个文本向量空间中寻找出 K 个最相似的文本，统计这些文本所属的类别，并用出现最多的类别号作为该未知文本的类别。其中，利用文本向量之间的余弦相似度[15]来度量文本之间的相似性。整个过程可以由算法 3 来描述：

算法 3： Classify(TC，Test Corpus，TSet， K)

输入: 训练词集 TC，测试集 Test Corpus，有效特征集 TSet，近邻个数 K

输出: 测试集中各文档的类别

描述:

```
Step 1: # 对算法 2 中获得的有效特征集 TSet，计算每个特征 t 的权重
        for 对于 TSet 中的每个特征 t:
            for TC 中的每个文本 d：
```

根据公式 TF-IDF-CF 计算 w(t, d)

Step 2: # 对测试集进行分类

for 对于每个测试集 Test Corpus 中的文档 file：

计算 file 中有效特征的权重

遍历整个训练空间，使用余弦相似度，计算该文档 file 和训练文档的相似度，找到 K 个最相邻的文档

从这 K 个最相邻的文档中，选择类别最多的作为该文档的类别

在算法 3 中，根据 TF-IDF-CF 公式对特征空间进行权重计算，是体现了每个特征词对特定类别的贡献程度，从而能够在自动计算简便的同时，得到更准确的文本分类效果。

5 实验及其分析

5.1 数据集

本文实验采用 NewsGroup[16]这个国际上通用的数据集。它是互联网用户在 Usenet 上张贴的 19997 条消息组成的。这些消息均匀分布在 20 个不同的新闻组中，每个新闻组约有 1000 条消息，每个新闻组对应着一个文本类别。本文采用完整的 20 个类别作为数据集，其中将 12000 个文本作为训练集，剩余的 7997 个文本作为测试集。

5.2 性能评估标准

评价分类器性能好坏的两个常用性能评价指标为召回率 r（recall）和准确率 p（precision），其分别定义为：

$$r = \frac{a}{a+c},\ if\ a\ +\ c\ >\ 0;\ otherwise\ r\ =\ 1,$$

$$p = \frac{a}{a+b},\ if\ a\ +\ b\ >\ 0;\ otherwise\ p\ =\ 1,$$

其中 a 表示被正确分到该类的文本个数，b 表示被误分到该类的文本个数，c 表示属于该类但被误分到其他类别的文本个数。

另一个常用的评估指标是 F-指标，它定义为：$F_{\beta}(r,p) = \frac{(\beta^2+1)pr}{\beta^2 p + r}$，其中参数β用来为召回率（r）和准确率（p）赋予不同的权重，当β取 1 时，准确率和召回率被赋予相同的权重。

本文中，使用准确率、召回率和 F-指标这三个指标来衡量分类器的分类性能。

5.3 实验结果及其分析

实验过程中使用到的参数设置如下：

全局低频特征百分比 z-low 为 1%；

全局高频特征百分比 z-high 为 1%；

DF 阈值 df-threshold 为 1%；

类别出现率 cf-rate 为 90%；

类别频率差异阈值 cf-threshold 为 0.0005；

权值公式中的调节系数 δ 为 0.01；

K 近邻算法中 K 为 500；

F-指标中的β为 1。

针对 NewsGroup 数据集，测试得到有 88336 个特征词，在经过全局特征提取和局部特征提取之后，总共约有 70000 个特征词，平均每篇文档包含 56 个有效特征词。可以看出，每个文本包含的特征词是大大减小了。

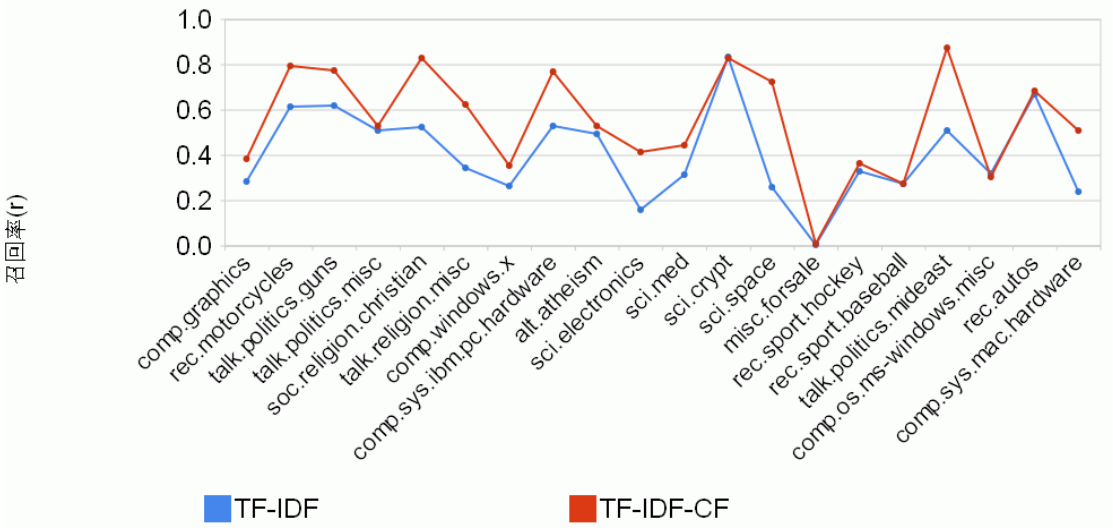

图 1 召回率比较

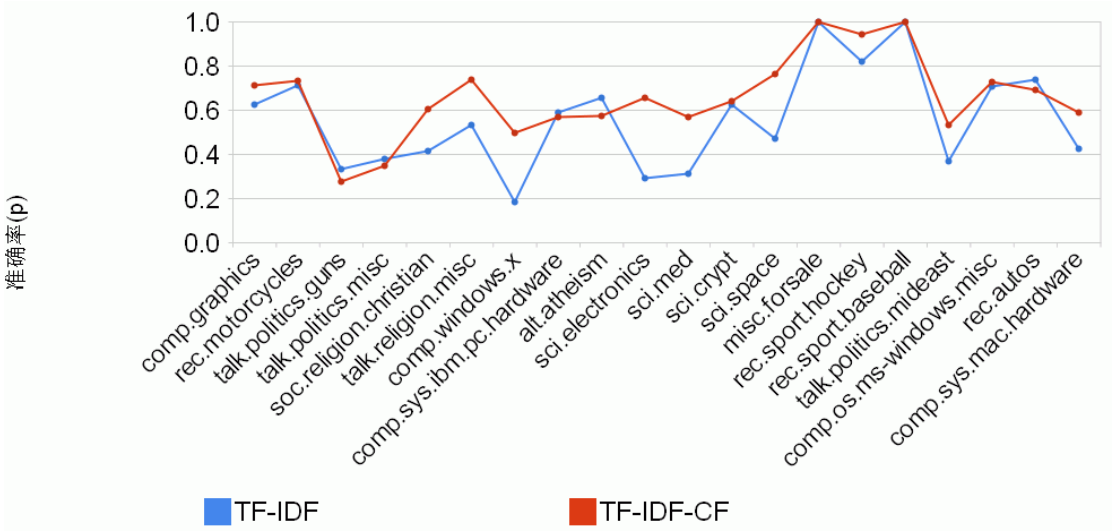

图 2 准确率比较

图 1 和图 2 分别显示在同一个特征空间中，使用本文提出的 TF-IDF-CF 和传统 TF-IDF 后，文本分类的召回率（r）和准确率（p）的不同结果。可以看到，在大多数文本类别中，都能取得较高的召回率和准确率，所以本文算法是优于 TF-IDF 方法的。

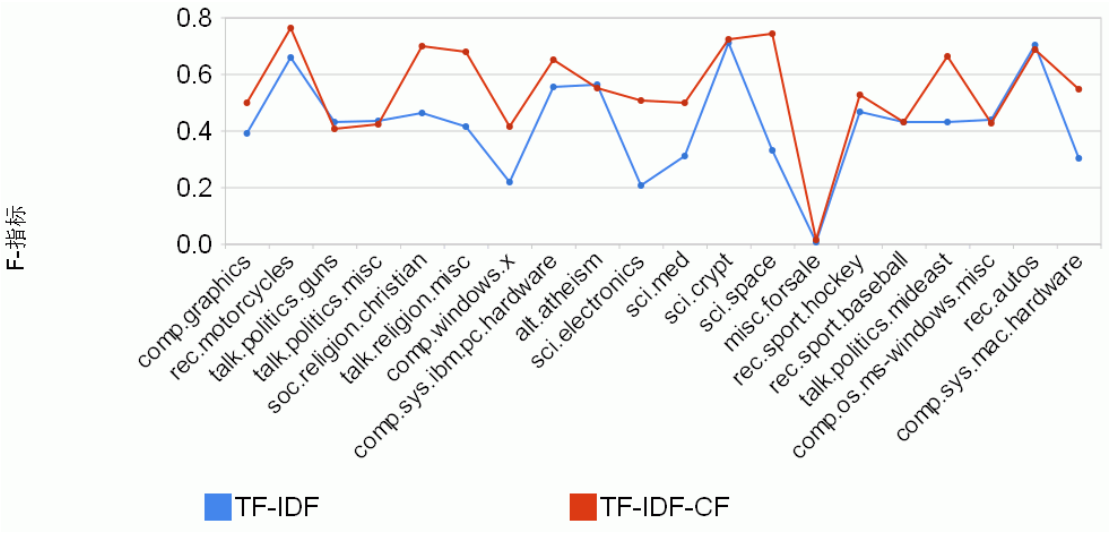

图 3 F-指标测试

图 3 显示了两种方法的综合评价，F-指标的测试结果。其中，β等于 1，即表示赋予召回率（r）和准确率（p）相同的权重。从图中也可以看到本文算法的综合评价也是高于传统方法的。

6 总结与展望

本文根据特征空间中特征本身的统计规律分别进行全局特征提取和类内局部特征提取。在得到有效特征之后，利用类别频率对传统的 TF-IDF 公式进行改进提出一种新的特征权重计算方法 TF-IDF-CF，在此基础上进行文本分类。从实验测试结果中看到，本文所采用的特征提取方法是有效的，根据改进的权重公式获得的分类准确度也是较高的。

参考文献:

[1] 李荣陆,王建会,陈晓云,等. 使用最大熵模型进行中文文本分类[J].计算机研究与发展,2005,42(1):94~101

[2] 王美方,刘培玉,朱振方. 基于 TFIDF 的特征选择方法. 计算机工程与设计,2007,12,Vol.28,No.23

[3] 冯长远,普杰信. Web 文本特征选择算法的研究. 计算机应用研究,2005,7,Vol.22,No.7:36~38

[4] 罗欣,夏德麟,晏蒲柳. 基于词频差异的特征选取及改进的 TF-IDF 公式. 计算机应用,2005,9,Vol.25,No.9

[11] Zipf Curves and Website Popularity. http://www.useit.com/alertbox/zipf.html

[5] 呼声波,刘希玉. 网页分类中特征提取方法的比较与改进. 山东师范大学学报(自然科学版),2008,9,Vol.23,No.3

[6] 褚力,张世永. 基于集成合并的文本特征提取方法.计算机应用与软件,2008,10,Vol.25,No.10

[7] 吴迪,张亚平,殷福亮,李明. 基于类别分布差异和 VPRS 特征选择的文本分类方法. 电子与信息学报,2007,12,Vol.29,No.12

[8] 徐燕,李锦涛,王斌,孙春明. 基于区分类别能力的高性能特征选择方法. 软件学报 Journal of Software,Vol.19,No.1,2008,1,pp.82-89.

[9] Qiang Wang,Yi Guan,XiaoLong Wang,Zhiming Xu. A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classification. IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.1A, January 2006.

[10] G.Zipf,Human Behavior and the Principle of Least-Effort (Cambridge, Mass, 1949; Addison-Wesley,1965);

[11] Xiaojin Zhu.CS838-1 Advanced NLP: Words, Zipf's Law, Miller's Monkeys. 2007 . http://pages.cs.wisc.edu/~jerryzhu/cs838/words.pdf

[12] 季燕江. Zipf 定律及其应用. http://www.qiji.cn/eprint/abs/4.html

[13] 韩筱璞. 网络信息搜索中的 Zipf 定律. http://www.qiji.cn/eprint/abs/840.html. [8]

[14] 张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类. 计算机工程. 2005, 4, Vol. 31, No. 8(171~173)

[15] Yiming Y. An evaluation of statistic approaches to text categorization[J]. Information Retrieval, 1999, 1(1/2):69-90.

[16] NewsGroup. 1999. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html