**在VSM中文本分类基于信息熵的文本加权方法**

**摘要**

文本加权方法已经被广泛的应用于信息检索和文本分类模型之中。在这篇文章中，在文本分类的背景下我们第一次研究集中先进的文本权重方法的局限性。鉴于类别明确的词语往往对于区分不同的类别有更好的作用，并且这些词语在这些类别中往往拥有更小的熵，所以我们就在探索一个词语区

分的能力它在一组类别中熵的关系。为此，我们提出了两个基于信息熵的文本权重方法（i.e tf·dc 和 tf·bdc）,这就意味着一个词语的区分能力取决于它在一个文集中的全局分布水平。为了演示我们提出的词语权重方法的效率，我们把它们和七种先进的方法分别在长文本集合短文本集上做了比较。我们的实验结果证明我们提出的方法要超过这些最先进的方法在用KNN和SVM在文本分类上面。

Ⅰ引言

随着日益增长的数字文件，文本分类（TC），一种可以自动将文档分类到预先设定的类别当中的技术，已经成为一种有效的技术来组织这些材料。为了完成一个文本分类任务，一个重要的一步是文档的表示。在向量空间模型（VSM）中，文档被表示为词语的向量来被分类器处理。既然不同的词语有不同的重要程度来表示文档的语义，那么给词语分配适当的权重的方法已经广泛的应用在文档的表示中来提高文档的识别率。虽然一些分类器（例如，SVM）可以自己学习权重，但是在文档的表示中，词语的权重是将词语在向量空间中映射到适当的位置。一个高质量的映射，可以帮助分类器更加有效的分类文档，从而让分类器表现得更加的好。

词语的权重在文本的识别中可以解释为测量一个词语在描述不同类别的效用。目前，词语的权重方法被广泛的分为无监督的和有监督的两种，这是根据他们是否利用训练文档的类别信息。大多数的无监督的方法都是从信息检索（IR）中派生的，比如说tf,tf·idf 和一些其他的变种。因为训练文档的类别标签是被忽视的，大多数的无监督学习的方法，例如 tf·idf 和BM25，实际上都是用来测量一个词语用来区分一个文档和其他文档的效用而不是区分一个类别和和其他类别的效用。因此，无监督的学习在文本分类中用来测量一个词语的区分能力是不够的。

相反，在监督方法里用训练文档的类别标签来对词语加权。然而，大多数的监督方法加权一个词语是基于一个词语在积极类别（PC,消极类NC）中的出现次数，这个是有下面这些缺陷的。第一，在多种情况下，积极类是一个单类，消极类是很多类的组合。这个词语的在消极类中统计（例如词语的出现和不出现）往往占有很大的权重在一些监督学习的方法中。因此，通过这些统计来反映词语的区分能力往往是不够有用的。第二，在大多数的监督学习中多个类往往直截了当的作为一个消极类。然而，一个词非常明显出现在消极类的不同类别中往往被忽视了。这导致了词语在类别区分上的信息缺失，所以词语间更加明确的区别不能够被反映。第三，这些加权词语的方法在一个给定的积极类基础上，然而测试文档是没有类别的标签。因此，这个挑战是用这些方法来正确的表示这些文档。

为了解决这局限，我们在这篇文章中提出了两个新的监督词语加权方法。通过我们的观察，一些词语只出现在某几个类别当中（称为类别明确的词语），相反，有些词语分散在很多的类别中（称为类别普通词语）在文库中。对于类别明确的词语能够更加好的指示类别，这些词语相比那些类别一般的词语更加的有效来区分不同的类别（也就是说有更好的区分能力）。在另一方面，一个词语越集中在一些少的类别中，那么他有更小的熵。因此，一个词的熵能够用来明确的表示一个词语在类别的分布以及反映了他在文本分类中的区分能力。例如“author”和”OPEC”是两个在Ruuters-21578的词语。图一展示了他们在八个路透社的类别中出现的频次我们可以发现“author”散列在每一个类别中有一个很大的熵。相反，“OPEC”主要集中在一个类别之中并且有一个很小的熵。明显的“OPEC”比“author”有更强的区分能力.基于这样的观察，我们提出了两种给予文本权重熵的方法来完成文本分类任务。我们工作的贡献总结如下。

我们首先来揭示和阐述这些在文本分类中最先进的词语加权方法的局限性。为了解决这些问题，我们基于熵的方法，这个加权的方法是基于这个词语在整个文集的类别中的分布度。由于没有在类别上使用消极类和积极类来分开，每个词语在每一个类中的出现都可以用tf·dc来很好的表示。再者，tf·dc 加权方法是和类无关的，因此可以恰当的表示测试文档虽然他们没有类别标签。

为了加强tf·dc在偏的类别分布上的表现，我们用一个词语在它相关类别的比例来代替在标准熵中的词语的分布。，由此，我们提出了一种新的基于熵的方法tf·bdc.

为了探索“采用熵的方法来测量一个词语的区分度是否比其他的方法更好”的答案。我们把提出的方法分别和7种最先进的方法在长文本和短文本集上进行比较。我们在实验结果上提出了详细的讨论并且我们还从很多的方面做出了总结。

**Ⅱ词语加权方法的研究**

一些词语加权方法的详细的阐述在8和13中已经给出。在这片文章中，我们主要揭示一下在文本分类中那些被广泛使用的词语加权方法的局限性。为了更加直观的展示，在这一部分中我们探索这些方法通过一些实际的例子。这些例子是从搜索语料库中获取的一部分。这个部分包括了从商业（Bus）,电脑（Com），教育（Edu）和工程（Eng）4个类别的1000篇文章。表一展示了在每个类别4个词语在文档中的频次。

1. 无监督的方法

大多数无监督的词语加权方法最开始是从信息检索中提出的。这种词语加权方法主要是靠词语出现的次数，但是忽略了训练文档的类别标签。一些被广泛使用的无监督方法包括bianry,tf,tf·idf以及其他的一些变种。

词语的频次（tf）在一个文档中是最直观的词语加权方法。目前，tf有像log(tf),log(tf+1)和log(tf)+1等变种，然而目前的研究表示这些变种之间没有显著的区别。这样，我们在这篇文章中只需要就原始的tf就可以了。 任何的tf方法可以用0或来代替，因此我们可以把它合并在binary方法中。通常，rf作为一个文档的局部加权并且通常是和其他的全局因子一起来对词语加权。

在信息检索中文档频率的倒数（idf）在信息检索中是被广泛使用的全局因子。不像无监督的信息检索任务，在执行信息检索任务之前训练文档的类别关系是已知的。在文本分类中，一个词语高频率的集中在一个类别中，那么这个是一个很好的分辨着，所以他因该被给一个很高的权重。然而，idf方法由于它的无监督的性质不能反映类别词语的强度。让我们来观察图一中的例子吧。

例子1：因为“IPO”的文档频次是23而“AMD”是20，那么idf方法应该分配给“AMD”比“IPO”更大的一个权重。然而，“IPO”只出现在Bus中，而“AMD”散列在Bus和Com中。作为Bus的一个好的分辨者，“IPO”直观上显然比”AMD”有更好的区分能力。因此，在文本分类中，idf在反应词语的分辨能力方面是不足的。

事实上tf·idf测量的是一个词语从一个文档区分其他文档的效用而不是区分一个类和其他类的效用。虽然有很多tf·idf的变种提出来像BM25和tff·tdf,这些方法和tf·idf有着同样的基础，因此在文本分类中他们和tf·idf有着相同的局限性。

Dumais用一个基于熵的因子来替换idf并且报告了这个比tf·idf更出色的新方法。然而，Dumais计算一个词语的熵是通过每一个文档得来的，这是一个计算量十分庞大的任务，尤其是在文集中有很多的文档的时候。除此之外，在这个方法中前面训练文档的类别信息是没有被考虑的。因此，相比于其他无监督的方法，在文本是被中这个方法不足以反映词语的区分能力。

1. 监督的方法

最近，一些学者已经利用训练文档的类别信息来对词语加权并且提出了一些监督的词语加权方法。大多数的监督学习方法是类别明确的，换句话说，每个词语在与它相关联的明确的类别中有一个权向量。例如，为了评估一个词语t的区分能力，每一个类别被轮流的当做一个积极类（PC）然后其他类被组合起来作为消极类。表格2展示了在和中包含和不包含t的文档的数量，并且N=a+b+c+d。那么t对于的区分能力是由这个统计表加权的。这些方法直观上来看，一个词语的区分能力越强，那么这个词语在积极和消极类上的分布月不同。为了和表一中的例子保持一致性，在这和例子中我们假设Bus类是积极类。

1）.基于特征选择的方案：目前，大多数的监督学习方法是基于特征选举取方法，这些方法使用不同的指标，例如chi-square(chi),gain ratio(gr),(Log-likelihoood Ratio)和eccd等。因为测试据报道和chi测试对于大样本是一致的，所以我们在我们的实验中只包含chi测试。这些方法的基本思想是一个词语和PC的关联越密切，那么就假定这个词语有更大的区分能力。

在8和13的代表，大多数特征选取加权词语的方法是基于统计表的，并且他们把一组类别分为PC和NC。然而，在多个类别中，PC是一单单一个类别，而NC是很多类的组合。这样可能造成一个很大的不平衡在PC和NC之间。因此，在表2中d项相比于其他三个可能很大，从而决定了特征选择方法的结果。然而d和a,c比在表明词语的区分能力方面是缺乏意义的。因此，这些特征选取方法在多类别因素上的反映词语的区分能力是缺乏有效性的。值得注意的，虽然eccd使用一个词语在类别上的熵作为加权的组合，但是这个方法依然是基于统计表并且包含d在另外的加权组合。

2).最新进的方法：为了避免b和d支配加权的结果，关联频率使用a和c的比例来评估一个词语的区分能力。然而rf有基于PC/NC分开方法的缺陷。一个词语在不同的NC类中的出现被组合在一起成为一个c。在NC中一个词语的区分信息，比如一个词语在多少个类中出现就丢失了。这样就导致了用这个方法反映的一个词语的区分能力就不那么合适了。让我们来考虑一个例子。

例子2.“AMD”和“private”在表1中的PC中有相同的出现次数（就是a），但是有不同的出现次数在NC中（也就是c）,“AMD”8次，“private”7次。有rf的方法计算，“AMD”比“private”有更小的能力。然而，“AMD”比“private”出现在更少的类别中.直观的，一个词语出现在更少的类别中那么它更能说明文档的类别。因此，“AMD”比“private”有更大的区分能力，这个和rf的结果相反。

虽然已经提出了一些rf的变种，像对数缩放的rf（叫做vrf）和概率的rf（叫做trr），这些变种因为和rf有着相同的基础，所以在文本分类中有着相同的局限。Iqf·qf·icf是最近提出来的另外一种方法，相比于rf一个词语出现的类别的个数，这个被叫做类别频次（cf）在Iqf·qf·icf中被列入考虑吧之中。然而，单单一个数字是不足以反映一个词语在类别中的分布的。比如一个词语分布在每一个类别中。让我们来看另外一个例子。、

例子3.在表一当中，因为“private”比“market”出现在更少的类别当中，Iqf·qf·icf方法分配给“private”更大的权重。然而，相比于他们的类别分布，“market”有更大的可能心出现在Bus中。因此用“market”来区分Bus相比“private”有更高的可能性。所以，“market”似乎比“private”有更好的区分能力，这个和Iqf·qf·icf的结果相反。

再者，一个基于统计学的信心的方法已经被提出来了在17中。虽然这个监督学习的方法比tf·idf和tf·gr更出色，但是它相比其他方法实现起来是复杂和困难的。

Ⅲ 基于熵的词语加权方法

1. 基于熵的方法的直觉

就像上面的讨论一样，除了出现在PC类中的词语，那些出现在NC类中类别明确的词语对于反映一个词语的区分度也是有用的。比如在表一中展示的励志一样，那个只出现在Bus中的词语“IPO”是Bus类的一个很好的指示器。因此，“IPO”比其他的词语有更好的区分能力。虽然“market”出现在很多的类中，但是大多数的“market”都集中在Bus中。因此，“market”相比于“AMD”，“private”有更大的确定性来指示Bus.所以，我们认为“market”比“AMD”和“private”有更大的区分能力。相似的，因为“private”是一个类别普遍的词语，那么“AMD”有更大的确定性来指示一个文档的类别。也就是时候，“AMD”比“private”有更大的区分能力。因此，我们可以得到如下结论：

结论1：一个文档中给定的两个词t1和t2,如果t1只分布在文集中很小的一些类别中，然而t2均匀的分布在很多的类别当中，那么t1比t2有更大的确定性来指示一个文档的类别；也就是说在文本分类中t1比t2有更大的区分能力。

根据结论1，我们假定一个词语的区分能力由全局分布度，也就是词语在类别的分布情况决定的。一个词语的集中性越高，那么它的区分能力越强。在另一方面，一个词语在类别中有更集中的分布，那么他有更小的熵。因此我们得到如下假设：

假设1：在一个文档中，给定两个词语t1和t2，如果在文集的类别中，t1比t2有更小的熵，我们假设在文本分类中t1比t2有更大的区分能力。

根据假设1，在文本分类中，一个词语的区分能力是由一组类的熵决定的。然而，在我们已知的，这个假设在当前的研究中还没有被证明。为了证明我们的假设，在这篇文章中我们提出了两种基于熵的方法。

1. 分布的集中度

既然一个更小的熵指示一个更高的集中度和更大的区分能力，我们定义我们的第一个监督学习方法是分布集中度（dc）：

，（1）

这里H(t)在文集的类别中词语t的熵。表示t在类中的频次这是根据训练文档的类别标签类别标签。f(t)表示t在所有类别中的频次总和。因为H,dc的值在被除之后标准化到[0,1]。这样保证了在同一个文档中词语是可以相互比较的。

就如表一展示的那样，dc的加权方法来反映这些词语的区分能力更加的合理和直观。因为搜索片段是十分的短小的，每一个词语大多数只出现了一次在每一个文档中。为了比较直观的和其他的词语比较，我们假设在表1中的值用表示。然而，这些假设只在这些例子中成立，在实践中是根据它的定义来计算的。

不像无监督的方法，我们提出的dc方法是一个监督的方法，它会利用训练文档的类别信息不像在[2]中计算熵要在每个文档中，我们计算熵是在每个类别中。因为一个文集中类别的数量是远小于文档的数量，因此在类别上计算熵的计算成本要比在文档上计算熵小得多。不像特征选取的方法，dc没有包含b和d的值，这样就避免了值主导了加权的结果以及减少了结果的有效性。虽然ig包含了熵的计算，但是ig加权一个词语t是通过测量在一个文集中把t设置为分离器后类别信息分布改变的信息来决定的，并且在ig中熵是用来量化设置分离器t后状态的信息。相反，我们加权一个词语是基于这个词语在类别的分布然后直接用熵来直接测量明确度。不像类比明确方法，dc是一个类别独立的方法，词语的在类别上的加权是基于词语的全局分布聚集度。也就是说，dc分配权重是不需要要求一个预先指定的PC。因此，dc可以恰当的表示文档虽然他们没有类别标签。

Ⅳ 总结以及未来的工作

正在这篇文章中，我们探索了几种最先进的方法的局限性并且提出了两种基于熵的加权方法tf·dc和tf·bdc。不想大多数的监督学习方法，把多类别的情况转化为二分法的情况，因为我们加权的方法是通过全局的分布集中度，这个是考虑了文集中的所有类别。我们的试验证明我们提出的方法在文本分类任务中比那些最先进的方法表现的更加出色。因此，在第一节中的答案就是，在文本分类中用熵来反映词语的区分能力比其他的方法更有效。此外，我们还有下面这些发现。首先，那些稀有的词语就像那些噪声词一样对于tf·idf是超负荷的，因此在KNN中需要用更多的邻居来减少噪声的影响来实现最优表现。第二，基于特征选择的方案在二分类别中比多类别表现更好。第三，考虑bc的因数，bdc在小类别上的分类效果显著提高。最后，在SVM中在文档中的一个词语的局部加强是更加的重要，并且在SVM中用tf的方法比那些不用tf的方法表现更加的好。

在将来，我们会证明我们提出的方法在其他的数据集和分类器中的有效性。