文本分类入门（番外篇）特征选择与特征权重计算的区别

Jasper 不求上进的咸鱼 2019-12-30

在文本分类的过程中，特征（也可以简单的理解为“词”）从人类能够理解的形式转换为计算机能够理解的形式时，实际上经过了两步骤的量化——特征选择阶段的重要程度量化和将具体文本转化为向量时的特征权重量化。初次接触文本分类的人很容易混淆这两个步骤使用的方法和各自的目的，因而我经常听到读者有类似“如何使用TF-IDF做特征选择”或者“卡方检验量化权重后每篇文章都一样”等等困惑。

文本分类本质上也是一个模式识别的问题，因此我想借用一个更直观的例子来说说特征选择和权重量化到底各自是什么东西，当然，一旦解释清楚，你马上就会觉得文本分类这东西实在白痴，实在没什么技术含量，你也就不会再继续看我的技术博客，不过我不担心，因为你已经踏上了更光明的道路（笑），我高兴还来不及。

想想通过指纹来识别一个人的身份，只看一个人的指纹，当然说不出他姓甚名谁，识别的过程实际上是比对的过程，要与已有的指纹库比较，找出相同的，或者说相似到一定程度的那一个。

首要的问题是，人的指纹太复杂，包含太多的位置和几何形状，要完全重现一个人的指纹，存储和计算都是大麻烦。因此第一步总是一个特征选择的问题，我们把全人类的指纹都统计一下，看看哪几个位置能够最好的区分不同的人。显然不同的位置效果很不一样，在有的位置上，我的指纹是是什么形状，其他人也大都是这个形状，这个位置就不具有区分度，或者说不具有表征性，或者说，对分类问题来说，它的重要程度低。这样的位置我们就倾向于在识别的时候根本不看它，不考虑它。

那怎么看谁重要谁不重要呢？这就依赖于具体的选择方法如何来量化重要程度，对卡方检验和信息增益这类方法来说，量化以后的得分越大的特征就越重要（也就是说，有可能有些方法，是得分越小的越重要）。

比如说你看10个位置，他们的重要程度分别是：

1 2 3 4 5 6 7 8 9 10
（20，5，10，20，30，15，4，3，7， 3）

显然第1，第3，4，5，6个位置比其他位置更重要，而相对的，第1个位置又比第3个位置更重要。

识别时，我们只在那些重要的位置上采样。当今的指纹识别系统，大都只用到人指纹的5个位置（惊讶么？只要5个位置的信息就可以区分60亿人），这5个位置就是经过特征选择过程而得以保留的系统特征集合。假设这个就是刚才的例子，那么该集合应该是：
（第1个位置，第3个位置，第4个位置，第5个位置，第6个位置）

当然，具体的第3个位置是指纹中的哪个位置你自己总得清楚。

确定了这5个位置之后，就可以把一个人的指纹映射到这个只有5个维度的空间中，我们就把他在5个位置上的几何形状分别转换成一个具体的值，这就是特征权重的计算。依据什么来转换，就是你选择的特征权重量化方法，在文本分类中，最常用的就是TF-IDF。

我想一定是“权重“这个词误导了所有人，让大家以为TF-IDF计算出的值代表的是特征的重要程度，其实完全不是。例如我们有一位男同学，他的指纹向量是：
（10，3，4，20，5）

你注意到他第1个位置的得分（10）比第3个位置的得分（3）高，那么能说第1个位置比第3个位置重要么？如果再有一位女同学，她的指纹向量是：
（10，20，4，20，5）

看看，第1个位置得分（10）又比第3个位置（20）低了，那这两个位置到底哪个更重要呢？答案是第1个位置更重要，但这不是在特征权重计算这一步体现出来的，而是在我们特征选择的时候就确定了，第1个位置比第3个位置更重要。

因此要记住，通过TF-IDF计算一个特征的权重时，该权重体现出的根本不是特征的重要程度！

那它代表什么？再看看两位同学的指纹，放到一起：

（10， 3，4，20，5）

（10，20，4，20，5）

在第三个位置上女同学的权重高于男同学，这不代表该女同学在指纹的这个位置上更“优秀“（毕竟，指纹还有什么优秀不优秀的分别么，笑），也不代表她的这个位置比男同学的这个位置更重要，3和20这两个得分，仅仅代表他们的”不同“。

在文本分类中也是如此，比如我们的系统特征集合只有两个词：

（经济，发展）

这两个词是使用卡方检验（特征选择）选出来的，有一篇文章的向量形式是

（2，5）

另一篇

（3，4）

这两个向量形式就是用TF-IDF算出来的，很容易看出两篇文章不是同一篇，为什么？因为他们的特征权重根本不一样，所以说权重代表的是差别，而不是优劣。想想你说“经济这个词在第二篇文章中得分高，因此它在第二篇文章中比在第一篇文章中更重要“，这句话代表什么意义呢？你自己都不知道吧（笑）。

所以，当再说起使用TF-IDF来计算特征权重时，最好把“权重“这个字眼忘掉，我们就把它说成计算得分好了（甚至”得分“也不太好，因为人总会不自觉的认为，得分高的就更重要），或者就仅仅说成是量化。

如此，你就再也不会拿TF-IDF去做特征选择了。

小Tips：为什么有的论文里确实使用了TF-IDF作特征选择呢？

严格说来并不是不可以，而且严格说来只要有一种方法能够从一堆特征中挑出少数的一些，它就可以叫做一种特征选择方法，就连“随机选取一部分“都算是一种，而且效果并没有差到惊人的地步哦！还是可以分对一大半的哦！所

以有的人就用TF-IDF的得分来把特征排排序，取得分最大的几个进入系统特征集合，效果也还行（毕竟，连随机选取效果也都还行），怎么说呢，他们愿意这么干就这么干吧。就像咱国家非得实行户口制度，这个制度说不出任何道理，也不见他带来任何好处，但不也没影响二十一世纪成为中国的世纪么，呵呵。

原文链接：
http://www.blogjava.net/zhenandaci/archive/2009/04/19/266388.html

用心去热爱编程和算法
就算题目数据再怎么水，
也不能降低自己的代码质量！

不求上进的咸鱼

阅读原文