

一种基于频次统计特性的兼类噪声消除方法¹⁾

蔡 巍 王永成 尹中航 李 伟

(上海交通大学计算机科学与工程系, 上海 200030)

摘要 本文着重研究了自动分类知识库中因为样本兼类而引起的存在于概念类频中的噪声, 提出了借助于统计特性来修正概念类频的算法。在进行理论分析的基础上, 本文讨论了算法的实现步骤, 并通过对新闻语料的分类实验, 检验了降噪效果。实验显示, 本方法可以减少兼类概念在知识库中的冗余次数, 提高自动分类系统的性能指标。

关键词 自动分类 知识库 降噪 自然语言处理

An Algorithm to Remove Multi-category Noise Based on the Statistics of Frequency

Cai Wei, Wang Yongcheng, Yin Zhonghang and Li Wei

(Department of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai 200030)

Abstract This paper studies the noisy in knowledge base for automatic classification, and presents the idea of text multi-category noise. By means of the statistical characteristic of knowledge base, a new algorithm is designed to deal with such noise. Experiment result indicates that our algorithm can obviously decrease the redundant appearing times of concepts of multi-category samples in knowledge base. The performance of automatic classification has been improved after revising.

Keywords knowledge base, reducing noise, natural language processing.

1 引 言

在基于统计的文本自动分类中, 表达文本主题概念与类别对应关系的知识库是一个非常重要的基础, 它直接决定了分类系统的准确程度, 因此, 人们提出了许多方法来减少知识库中的噪声。常用的方法是在标引和检索中减少非信息词 (non-informative words), 以此来增加知识的准确性和减少计算量^[1,2]。例如 Wilbur WJ 和 Sirotkin K 使用的 StopList 方法^[3], Yiming Yang 使用的 LLSF Mapping 方法^[4,5],

Qian Diao 使用的概念抽象法^[6], 以及 Lewis 和 Ringuette 使用的特征选择法^[7]等。这些方法对改善系统的性能有一定的作用, 但是忽视了这样一个事实: 由于文本的多主题性, 越来越多的样本呈现出多类别的特性, 即样本中存在着大量的兼类现象。利用这些兼类样本来构造知识库, 将会引入许多错误的概念类别对应关系。经过对知识库中各种噪声成分的分析, 本文给出了兼类噪声的概念, 并提出了一种借助于知识库的频次统计特性来减少兼类噪声的新算法。在给出了该算法的理论基础后, 本文详细地介绍了实现的具体步骤。我们选择了“中国资讯

收稿日期: 2003 年 9 月 3 日

作者简介: 蔡巍, 男, 1970 年生, 研究方向: 网络信息智能处理。王永成, 男, 教授, 博士生导师。尹中航, 男, 1968 年生, 博士。李伟, 男, 1976 年生, 硕士。

1) 该项目受国家自然科学基金资助 (批准号: 60082003)。

行”提供的20 000篇新闻语料作为训练样本,构建了概念类别对应关系表,并对该表进行了降噪处理。通过基于无修正 VMS 的自动分类实验,显示它可以显著地减少兼类概念在知识库中的出现频次,最多达到 16.57%。自动分类系统的准确率也有所提高,其中封闭测试增加了 5.2 个百分点,开放测试增加了 2.7 个百分点。与其他方法不同的是,本方法没有减少概念的绝对数量,因此对系统的分出率没有影响。因为在原理上独立于其他方法,因此可以较好地实现与其他方法的集成。

2 兼类噪声

一般情况下,基于统计的自动分类方法是通过训练样本来构建分类所需的知识库。图 1 给出了对这一过程的描述。

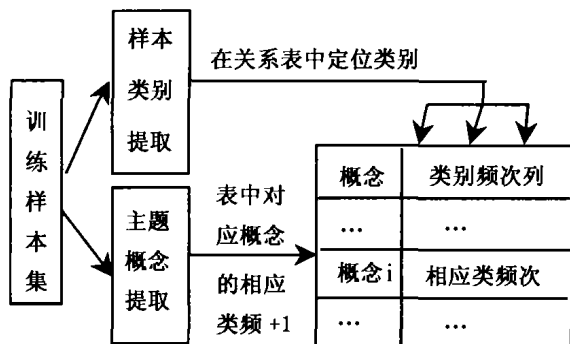


图 1 知识库建立过程示意图

通过对实际建库过程的分析,我们发现,除了样本不足、训练样本选择不当和主题标引不准确等引起的传统意义上的噪声外,在知识库中还存在着一种新的噪声——兼类噪声。该种噪声可定义为:一个多主题的训练样本(兼类样本)对应多个类别时引起的主题概念与类别之间错误的对应关系。

多主题性是文本的一个自然属性,对训练样本而言,意味着一篇文本同时对应着多个类别。当按照传统的方法,将该文本的一个主题概念简单地与多个兼类类别相对应时,常常会引入错误的对应关系。例如,有一篇文本, $D=(\text{计算机}, \text{社会发展})$,同时属于电脑类和社会科学类。当把这两个概念都简单地与这两个类别对应起来时,就会出现“社会发展”对应于电脑类,而“计算机”也对应着社会发展类的错误关系。

为了定量地检验兼类噪声,我们利用“中国资讯行”提供的 20 000 个新闻样本做了一个实验。这些

样本已经由该行的手工分类员进行了类别标注,每一个样本对应着 3 层共 196 个小类中的一个或几个类别。实验方法是用程序对每一个样本自动地进行结构识别(提取出对应的全部类别名)、类别纠错(自动地纠正手工分类时产生的类名错误)、类号转换(转换为标准类号形式)和类数统计。

表 1 是得到的结果。从中可以发现,有 $33.88\% + 3.4\% = 37.28\%$ 的样本具有两个以上的类号。按照每个样本具有 8 个主题概念来计算,并假设这些概念不重复,则在知识库中,共有 $20\,000 \times 8 = 160\,000$ 个主题概念。这些概念在所有类别中出现的总频次为 $12\,544 \times 8 + 6\,776 \times 8 \times 2 + 680 \times 8 \times 3 = 225\,088$ 。有 $6\,776 \times 8 + 680 \times 8 \times 2 = 65\,088$ 个概念可能对应着不正确的类号,占总频次的 28.92% 。这样大的比例(尽管不一定全部是噪声)将对分类精度产生影响,消除这样的噪声对提高分类精度具有积极的意义。

表 1 样本兼类情况表

样本类号数量	样本数量	百分比	兼类错误的概念频次(可能)
只有一个类号	12 544	62.72%	0
只有两个类号	6 776	33.88%	54 208
只有三个类号	680	3.4%	10 880

3 理论基础

我们的问题可以描述为:在将兼类训练样本中的一个主题概念放入知识库时,判定该概念属于兼类类别中的哪一个类别。我们依据 Bayes 最小错误决策方法来解决此问题。

设在给定的类别体系中共有 M 个类,记为 $C = \{c_1, c_2, \dots, c_i, \dots, c_M\}$ 。每类的先验概率为 $P(c_i)$, $i = 1, 2, \dots, M$ 。假设各个类别是平衡的,即 $P(c_i) = P(c)$ 。

对于一个概念 x ,其归于 c_i 类的类条件概率是 $P(x | c_i)$ 。在假设训练样本足够多的情况下,如果 x 在知识库中共出现了 n 次,并且在 c_i 类出现了 n_i 次,则有:

$$P(x|c_i) = n_i / n \tag{1}$$

根据 Bayes 定理,可得到 c_i 类的后验概率 $P(c_i|x)$ 是:

$$P(c_i/x) = \frac{P(x/c_i) \cdot P(c_i)}{P(x)} \tag{2}$$

按照最大后验概率判决准则,若 $P(c_i|x) > P(c_j|x), i = 1, 2, \dots, M, j = 1, 2, \dots, M$, 则有 $x \in c_i$ 。结合式(1)和式(2),此准则可以进一步表示为:

若有 $n_i > n_j, i = 1, 2, \dots, M, j = 1, 2, \dots, M$, 则 $x \in c_i$ 。

可以证明,在所有的归属度计算方法中,通过上述方法所得到的概念归属错误率或风险是最小的。

考虑到在非兼类的情况下一个概念可以对应着多个类别,为把这种情况与兼类区别开来,应该设定一个概率比例阈值。在构建知识库时,只有大于该阈值,才认为兼类样本中的概念确实只对应着兼类中的一个类别。

4 算 法

首先用传统的方法构造出概念类号关系表——源知识库,并由源知识库复制出一个新的知识库。以源知识库中各训练样本的概念在各个类别中出现的频次为依据,对每一个兼类样本中的每一个概念,检查在对应类别中的出现频次。如果出现的频次差别较大(倍数大于阈值),则认为此概念在频次较小的类别中因为样本的兼类而产生了一次噪声。因此将新知识库中此概念在此类别中的频次减 1,同时保持在其他类别中的频次不变。以此类推,当将所有的兼类样本的所有概念都处理完毕后,即可以认为消除了大部分兼类噪声。算法 1 给出了消除兼类噪声的详细步骤。

算法 1:消除兼类噪声流程

/算法开始:

按正常方式生成源概念类号表 Tab1;

由 Table1 复制出目标概念类号表 Tab2;

BEGIN

For (训练样本集中的每一个 Sam)

提取 Sam 对应的全部类号 CN (i), i = 1,2, ..., CM;

If (Sam 的类号数 CM> 1)

For (Sam 中的每一个概念 Con)

在 Tab1 和 Tab2 中找到 Con 对应记录 Rec1 和 Rec2;

在 Rec1 和 Rec2 中找到 CN (i)所对应的频次 -
Freq (i), i = 1,2, ..., CM;
对 Freq (i)从大到小进行排序,形成排序后的 -
Freq (j), j = 1,2, ..., CM;
For (j1 = 1 to CM - 1, j2 = j1 + 1 to CM)
If (Freq (j1) / Freq (j2)> Threshold)
For (j2 = j1 + 1 to CM)
Freq (j2) = Freq (j2)- 1;
把 Freq (j2)放入 Rec2 的相应位置;

END

用 Tab2 作为新的知识库表。

算法结束/

从上面的算法中可以发现,阈值(Threshold)的选取是决定噪声消除的关键因素。由于与此阈值相关联的不确定因素较多,无法用纯数学的方法确定最优值(即使能够确定,实际效果也不一定好),因此我们利用经验值方法来选取,即做出分类性能随阈值变化的曲线,从中找出最佳值。

下面再以能够产生兼类噪声的训练样本 D = (计算机,社会发展)为例,对该算法进行简单的说明。

设表 2 是由足够多的训练样本(包括 D)产生的部分概念类号关系表。其中概念“计算机”在电脑类中共出现了 100 次,在社会科学中则出现了 1 次,频次相差较大(100 倍)。概念“社会发展”在电脑类中共出现了 3 次,在社会科学中则出现了 200 次,频次相差仍然较大(为 66 倍),我们可以将阈值设为 10。由于训练样本足够多,因此可以认为“计算机”在电脑类中出现的 100 次反映了它与电脑类的正确的对应关系,故此频次保持不变。而它在社会科学中出现的 1 次则是由此样本兼类引起的(频次倍数大于 10),视为 D 在知识库中造成了一次兼类沙场,应当把此次数减 1。同样,也应当将“社会发展”在电脑类中的出现次数减 1,以便消除因样本 D 兼类而引入的误差。本次噪声消除后的类号关系表如表 3 所示。

表 2 消除噪声前的概念类号关系表

类别 \ 概念	...	电脑类	社会科学类	...
...
计算机		100	1	
社会发展		3	200	
...

表3 消除噪声后的概念类号关系表

类别 概念	...	电脑类	社会科学类	...
...
计算机		100	0	
社会发展		2	200	
...

5 实验与分析

5.1 实验方法

当样本较多时,知识库将变得非常庞大,无法用手工的方法来检验兼类噪声的消除情况,因此我们采用自动分类系统来实现目的。具体作法是:对降噪前后的两个知识库,用同一个自动分类系统对同一批样本进行自动分类,通过比较分类结果的差异来检验降噪效果。

5.2 实验环境

我们采用“中国资讯行”提供的 20 000 个新闻样本作为训练样本,另外利用 2000 个新闻样本作为开放测试样本,2000 个新闻样本作为封闭测试样本。同样地,这些样本分布在 3 层 196 个类别中。我们采用的主题标引系统是由上海交大近期开发的 ASI 系统。利用此系统,对 20 000 个训练样本进行处理,每个样本最多提取 8 个主题概念,并利用这些概念构造了原始的概念类号关系表。

分类方法采用的是未修正的向量空间模型(VSM),并利用向量内积作为相似性度量指标。分类的性能指标定义如下:

分出率(Recall)

= 分出的测试样本数 / 全部测试样本数

分准率(Precision)

= 被正确分类的测试样本数 / 分出的测试样本数

5.3 实验结果

按照上面的算法,通过设定不同的概念比例阈值,依据原始知识库和降噪后的知识库进行了封闭分类和开放分类,分类结果见表 4。

表4 试验结果

阈值	减噪次数	分出率		分准率	
		封闭测试	开放测试	封闭测试	开放测试
3	37302	90.7%	89.4%	72.8%	69.2%
5	15742	92.2%	90.0%	74.2%	68.4%
7	11001	92.2%	89.8%	74.6%	68.1%
9	6216	91.8%	89.8%	74.2%	67.9%
11	1729	91.8%	90.0%	70.1%	67.9%
13	280	91.4%	89.6%	69.4%	66.5%
原始	0	91.4%	89.6%	69.4%	66.5%

5.4 分析

从表 4 可以发现,不同的阈值所减少的概念频次是不同的,它与阈值的大小成反比,即阈值越大,减少的频次越小,如图 2 所示。这是因为当阈值增大时,在原始知识库中满足噪声条件的概念减少,因此能够去除的概念频次也减小。在本实验中,当此值为 3 到 5 时,去除的概念频次最多(当选择 3 时,可去除 37302 次概念),占概念出现总频次 225 088 的 16.57%。

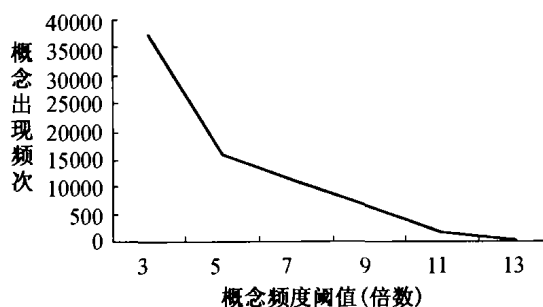


图2 减少概念次数随阈值变化情况

不同的阈值可以不同程度地改善分类的准确率,如图 3 和图 4 所示。对封闭测试而言,此阈值选择为 7 左右时改善的幅度最大,达到了 74.6% (69.4% + 5.2%);而对开放测试而言,此阈值为 3 时效果最好,达到了 69.2% (66.5% + 2.7%)。也就是说,封闭测试和开放测试的最佳阈值点稍有不同。出现这一现象的原因是我们在此论文中只考虑了兼类噪声,没有研究其他可能的噪声。而正是其他一些噪声(例如样本不足或不当引起的噪声),继续影响着概念类别对应关系的准确性,并造成了最佳阈

值点的“漂移”。

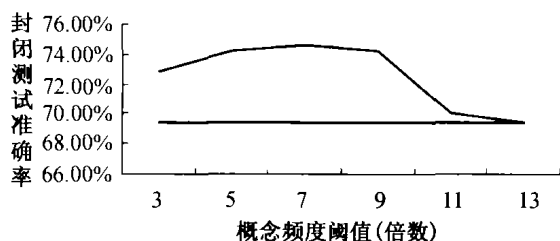


图3 封闭测试准确率随阈值变化示意图

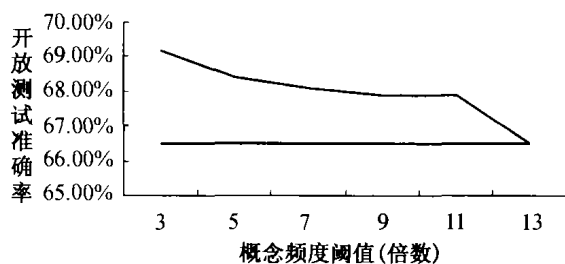


图4 开放测试准确率随阈值变化示意图

从表4中还可以发现,对不同的阈值,系统的分出率保持基本不变。这是因为本方法只是减少了可能产生噪声的兼类概念重复出现次数,而没有减少概念的绝对数量,这也是本方法与其他降噪方法的不同之处。

6 结 论

总结以上的讨论,我们可以得到以下几个结论:

(1)由于样本的多主题性,在训练样本集中存在着大量的兼类样本,如在我们的样本集中,兼类样本比例达到了37.28%。利用这些样本构造知识库时,不可避免地产生错误的概念类别对应关系(可能的错误对应关系在最坏的情况下可达到28.92%),这些错误关系将在一定程度上影响自动分类系统的准确性,因此有必要研究消除这些噪声的方法。

(2)以原有知识库的频次统计特性为依据,按照最小错误准则,对兼类样本中的概念进行消除,可以较大比例地减少这些概念在知识库中的出现频次,

从而达到减少错误的概念类号对应关系的目的。在我们的实验中,减少的频次比例最多可以达到全部概念出现总频次的16.57%。

(3)经过对原始知识库的兼类降噪处理,系统的性能指标有了一定的改善。其中,封闭测试的准确率最多增加了5.2个百分点,开放测试的准确率最多增加了2.7个百分点。

(4)本方法对概念的绝对数量没有影响,因此对系统的分出率没有影响。由于其他降噪方法主要考虑的是如何减少不重要的概念,因此本方法在理论上与其他降噪方法是不相同的,可以与其他方法实现原理独立的集成,从而充分发挥各种方法的优点。

下一步的研究将集中在以下三个方向上:一是研究在类别先验概率不为常数的情况下本方法的降噪能力,二是比较本方法与其他方法在降噪效果上的区别,三是实现本方法与其他各种方法的集成。

参 考 文 献

- 1 Salton G. Automatic Text Processing: The Transaction, Analysis, and Retrieval of Information by Computer. Pennsylvania: Addison-Wesley, Reading, 1989
- 2 Van Rijsbergen CJ. Information Retrieval, 2nd ed. London, England: Butterworths, 1979
- 3 Wilbur WJ, Sirotkin K. The automatic identification of stop words. Journal of Information Science, 1992, 18, 45 ~ 55
- 4 Yiming Yang. Noise Reduction in a Statistical Approach to Text Categorization. SIGIR, 1995
- 5 Yiming Yang, Wilbur WJ, Using corpus statistics to remove redundant words in text categorization. Journal of the American Society for Information Science (JASIS), 1996
- 6 Qian Diao. Research and Development of Knowledge Based Automatic Classification System for Chinese Information: [doctoral paper]. Shanghai Jiaotong University, 2000
- 7 D.D. Lewis and M. Ringuette. Comparison of two learning algorithms for text categorization. In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994

(责任编辑 许增祺)