维普资讯 http://www.cqvip.com

中 文 信 息 学 报
第18卷 第4期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.4

文章编号:1003-0077(2004)04-0009-06

基于规则的自动分类在文本分类中的应用*

李渝勤[1],孙丽华[2]

(1.北京信息工程学院,北京 100101; 2.TRS信息技术有限公司,北京 100101)

摘要:文本自动分类是指将文本按一定的策略归于一个或多个类别中的应用技术。本文首先介绍三种基于统计的自动分类技术(k近邻分类器、支持向量机分类器和朴素贝叶斯分类器),剖析了基于统计的自动分类的优势及不足。基于统计的自动分类的不足主要表现为:当类别之间分类特征的交叉变大时,分类精度呈下降趋势,在多层分类的情况下,此局限尤为突出。针对此局限性,为了提高自动分类的精度,我们引入了基于规则的自动分类来对其进行改进和扩充,并整合两种自动分类技术的优点,设计出了混合分类器系统,从而获得了比较理想的分类效果。

关键词:计算机应用;中文信息处理;文本挖掘;文本分类;规则分类

中图分类号:TP391 **文献标识码**:A

Rule-based Automatic Category Application on Text Category

LI Yu-qin[1], SUN Li-hua[2]

(1. Beijing Information Technology Institute, Beijing 100101, China;
2. TRS Infromaton Technology Limited Company, Beijing 100101, China)

Abstract: The technique of text automatic category is to classify texts into one or more classes according to some strategy. This paper firstly reports three kinds of technique of text automatic category based on statistic (k nearest neighbor , support vector machine and naïve bayes), and analyses their advantages and disadvantages. The weakness of statistic-based automatic category is the category precision decrease while the character intersect within classes increase, especially in the case of multi-layers classifying. In order to improve statistic-based automatic category performance, rule-based automatic category is used. we combine statistic-based category with rule-based classifying method , design and realize a system of mixing category lastly, which has and has had very good performance in category.

Key words: computer application; Chinese information processing; text mining; text category; rule-based classifying

1 前言

随着现代社会的进步,各种各样信息的迅猛发展,尤其是Internet网络资源的快速发展,使人类社会面临着日益严重的信息挑战。人们不仅重视信息的有效性,而且更加关注信息获取的经济性。如何便捷地获取信息,如何高效地应用信息,已经成为现代信息技术的研究热点。文本自动分类等文本挖掘技术就是在各种信息量异常庞大、信息载体纷繁复杂瞬息万变的形势下,应运而生的一整套的在各种文本载体中发现信息、处理信息的最佳方案,也是人们更加经济地获得有效信息的途径。

* **收稿日期**:2003-11-05
作者简介:李渝勤(1963—),女,高级工程师,主要研究方向为计算语言学.

维普资讯 http://www.cqvip.com

目前,国内外在文本自动分类方面的研究主要是基于统计方法并取得了可喜成果[1]。其主要算法有:k 近邻法(KNN),朴素贝叶斯法(naïve bayes),贝叶斯网络,神经网络算法(如:BP 算法等),支持向量机(Support Vector Machine),EM 算法, SOM 算法[2,3]。其中 SVM 分类器和 kNN 分类器是目前分类性能最好的两种分类器。本文分别介绍 kNN 方法,SVM 方法及贝叶斯方法,分析其优势,指出其不足,从而引进基于规则的自动分类方法,设计了一个改进方案,然后运用于实际系统,使自动分类的精度更高。

2 基于统计的自动分类方法

2.1 k 近邻法

kNN 方法是传统的模式识别算法。对于一个测试文本,计算它与训练样本集中每个文本的文本相似度[4],依文本相似度找出 k 个最相似的训练文本,然后在此基础上给每一个文本类打分,按分值进行排序,依分值指定测试文本的类别。为了分类合理,可以选定一个阈值。形式化表示为:

$$f(d_x, c_i) = \begin{cases} 1, & if \sum_{d_j \in kNNDoc} sim(d_x, d_j) \cdot g(d_j, c_i) - b \geqslant 0 \\ 0 & 其它 \end{cases} \tag{1}$$

2.2 支持向量机(SVM)

SVM 基本思想见图 1。

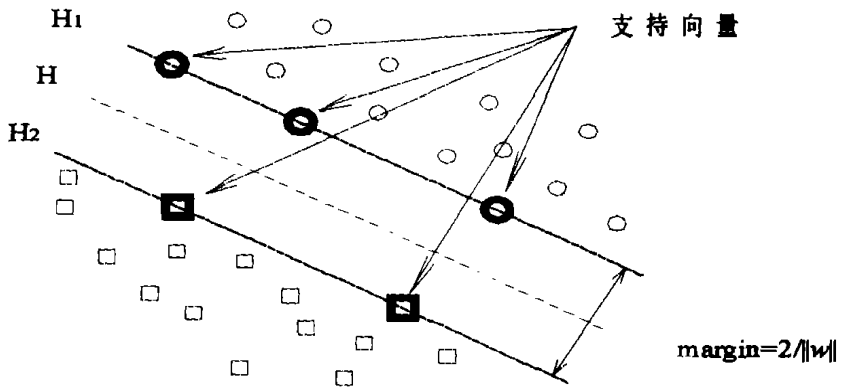

图 1 SVM 基本原理

利用 Lagrange 优化方法可将解 w 转化为如下二次优化问题

$$\text{minimize: } W(\boldsymbol{\alpha}) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

$$\text{subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\forall_i^n : \alpha_i \geqslant 0$$

α_i 为每个样本对应的 Lagrange 乘子。该问题存在唯一解,且解中只有一部分(通常是少部分)α_i 不为零,对应的样本就是支持向量。最后得到分类指示函数,它实际是只对支持向量求和,这就是支持向量机[7]。

$$f(\boldsymbol{x}) = \text{sgn}\{(\boldsymbol{w} \cdot \boldsymbol{x}) + b\} = \text{sgn}\Big\{\sum_{i=1}^{n} \alpha_i^* y_i (\boldsymbol{x}_i \cdot \boldsymbol{x}) + b^*\Big\} \tag{2}$$

2.3 朴素贝叶斯分类器

朴素贝叶斯分类器假设特征对于给定类的影响独立于其它特征。对文本分类来说,它假

维普资讯 http://www.cqvip.com

设各个单词之间两两独立[8]。设训练样本集分为 k 类，记为 $C=\{C_1,C_2,\cdots,C_k\}$，则每个类 C_i 的先验概率为 $P(C_i)$，其值为 C_i 类的样本数除以训练集总样本数 n。对于新样本 d，其属于 C_i 类的条件概率是 $P(d|C_i)$，C_i 类的后验概率为 $P(C_i|d)$[9]。

$$P(C_i \mid d)=\frac{P(d \mid C_i)P(C_i)}{P(d)} \propto P(d \mid C_i)P(C_i) \tag{3}$$

2.4 基于统计的自动分类的优势与不足

以上几种基于统计的自动分类方法，根据训练语料，得到各类别的模板，进而通过模板进行分类。其优点是训练简单方便，一般情况下分类精度高；其缺点主要有两点：1）当类别之间交叉现象比较严重时（两类之间的特征重复较多），分类器的精度会大大下降，尤其是在多层分类中，有些子类之间的特征交叉更为严重，因此在大类别基本正确的情况下，子类的分类精度却大大下降。例如，在对经济，历史，体育，科技，医药卫生等类别进行测试时，测试结果表明体育类的分类效果是最好的，精度达到99%左右，主要原因就是体育类的特征与其他类的交叉非常小；而医药卫生和科技的精度较低，都在90%以下，主要原因是这两个类特征有交叉与其他类之间也有交叉。2）对训练语料的数量与质量均有较严要求。如果语料不全面，代表性不强，则会直接影响自动分类的精度。每类语料要涵盖该类所涉及的所有方面，例如某类的某个方面的语料准备得不全，那么分类器在分类时，可能会分错。准备语料一定要确保语料所属类别的准确性，即某类的训练语料一定要属于该类，不能把不属于该类的语料标为此类，否则会直接影响分类器的分类效果。

3 基于规则的自动分类

3.1 基于规则的自动分类

基于规则的自动分类的基本思想：用户直接为每个类目确定分类规则形成类别模板，规则分类器依据类别模板统计测试样本中满足的规则条数及规则出现的次数信息，同时利用规则在测试文本结构中的位置信息，来衡量测试样本所属的类别。

类别模板的形成过程：每个类别模板可以由很多条规则组成，在类别模板中每一行信息代表一条规则，每条规则可以由多个项组成，每一项都是该类的关键词。规则支持“与”“或”“非”“异或”等逻辑运算。同时支持字段信息。逻辑运算顺序为：

高　　（　）括号
　　=　等于
　　*　与
　　^　异或
　　-　非
低.　　+　或

例1：一条规则可写为：（标题=计算机+作者=张三）*正文=电脑

意义：文章的标题中出现“计算机”或作者字段中出现张三，同时正文中出现“电脑”则该条规则成立。

例2：规则为：正文=（汽车-自行车）

意义：正文中出现“汽车”但不能出现“自行车”则该条规则成立。

基于规则的自动分类的关键是制定的规则是否完备，准确，有代表性（代表整个类别信息）。规则的构造过程采取三方面的途径：

维普资讯 http://www.cqvip.com

1. 与分类领域专家合作，由专家来制定规则，保证规则的科学性。例如新华社新闻分类项目中，与新华社新闻分类专家合作共同制定分类规则。

2. 通过各种分类主题词典来获得类目的主题词信息进而制定规则。可以参考的一些分类词典如《现代汉语分类词典》、《中国分类主题词表》等。

3. 通过统计算法来提取类别的关键词，再通过人工进行筛选之后制定规则。

通过以上三种方法，TRS 系统库中建立了地区分类、体育分类、行业经济分类等一系列子规则库，总类目数达到上千个，规则近万条，随着用户的需求规则库不断的扩充。

基于规则的自动分类与基于统计的自动分类的区别是基于规则的分类不需要提供训练语料。基于规则的分类的优势是分类精度高，对规则模板可以随时修改，灵活方便；缺点是规则要全面，要有代表性，否则就会直接影响分类器的性能，当类目规模增大时，需要确定的规则数量增多，而且规则的维护比较麻烦。

3.2 基于规则的自动分类对基于统计的自动分类的补充

通过上面对基于统计的自动分类与基于规则的自动分类的分析，可以采用基于规则的分类来对基于统计的分类进行有效的补充和改进。

当类别之间特征交叉大时分类精度下降，尤其对多层分类，子层的类别间交叉相对会更大，自动分类器对子层分类准确率会降低，进而影响整体的分类性能。针对这种情况可以采用规则的分类来替代统计的分类对交叉大的层次进行分类，提高分类精度；具体的策略是：对于整个分类体系中，第一层的类别一般主要采用自动分类，因为第一层的每个类都包含很多方面的内容，很难用规则写的全面。对各个大类下面类别之间交叉现象很严重的子类，可对这样的子类采用基于规则的自动分类，如果该层子类仅某一个类别与其他的类别交叉较大，则可以对该类采用基于规则的分类，其余的类别采用基于统计的分类。分类器分类过程是：先通过第一层的分类器进行分类，如果属于第一层的某个类，再用该类的下一层分类器进行分类。如果属于第二层的某个类，且该子类有第三层，则再用第二层的该子类(第三层的类别)分类器进行分类，一直分到叶子节点。如果某层采用的是基于统计的自动分类，就用统计分类器；如果采用的是基于规则的自动分类，就用规则分类器；如果两者都采用，则就分别采用这两种分类器对测试文本进行分类，对得到的两组分类结果根据需要取交集或并集等。例如：股市，汇市，利率这三个类之间的交叉较大，基于统计的自动分类效果不是很理想，因此可采用基于规则的分类来进行自动分类，见测试 1。

基于统计的自动分类对训练语料的数量和质量的要求高，当一些类目没有充足的训练语料或语料的质量不高时，分类效果不好，可以采用基于规则的分类来代替基于统计的分类。具体哪些类目采用规则方法代替统计方法，要根据实际的类目及语料准备的情况而定。

此外，对于类目特征词比较明显，特征词表较小，规则的制定比较容易时，也可采用基于规则的分类。例如体育下子类乒乓球，篮球，足球等，这些子类的特征词表较小，规则制定也比较容易，因此可采用规则分类。在应用中，也会遇到某层的一些类没有语料，无法采用基于统计的分类方法，也可以采用规则的分类的方法实现。

在实际的分类中，有的用户强调准确率，有的用户强调召回率。如果基于统计的自动分类不是很理想的情况，可以采用基于统计的分类和基于规则的分类同时进行分类，对两种分类方法的分类结果进行求并集或求交集的处理。如果强调准确率，则对两种方法的分类结果进行求交集处理；如果强调召回率，则对两种方法的分类结果进行求并集处理。例如，在网络信息雷达系统中，用户要求从网上抓取到网页，通过分类器进行分类，然后按照类别把网页信息发

维普资讯 http://www.cqvip.com

布出去,该系统强调分类的准确率,对分类的召回率要求不高,对于没有分出类别的网页还设了一个类别"其它",这样没有分出来的都归到其它类,针对这种情况就可以对两种方法的分类结果的求交集处理,来提高分类网页的准确率。

4 文本自动分类系统

经过大量的理论的研究与实践的探索,设计和实现了 TRS 文本自动分类系统。该系统经过长期的测试与优化,性能稳定,分类精度高。下面对该系统的设计与性能进行介绍。

4.1 系统设计

TRS 文本分类系统主要采用混合分类器引擎的设计方案,即基于规则的分类与基于规则的分类相结合的设计思想,统计分类器采用支持向量机(svm)与 K 近邻法(knn)分类方法,规则分类器主要采用统计算法与语言学知识相结合建立规则模板进行分类。具体某层或某些类,用基于规则的分类还是用基于统计的分类,用户可以根据需要随机的组织进行设定,灵活方便。判断类别之间的交叉情况主要采取两种措施,一种为人工来实现判断类别的交叉情况,进而决定是否使用基于统计的自动分类;另一种为通过统计每类的特征词,来比较类别之间的特征词的重合程度,来判断类别的交叉情况。

考虑到新闻语料的实时变化性特点,用以前的分类模板来分类当前的文本,会存在大量错分和漏分的问题,针对这种情况设计并实现了反馈学习机制。即根据用户的反馈信息,分类器自动对分类模板进行优化和完善。同时,从用户的实际使用角度,本系统增加补充训练机制。对用户而言,类别的要求是随时可能变化的,当需要增加新类时,如果把以前的类都重新再训练一次,会给用户带来使用上的不便和时间上的浪费。这样,可以方便用户,节省用户的时间。

4.2 测试

测试 1:

数据来源:新浪网站,搜狐网站,新华网,人民网

类别:股市　利率　汇市

训练文本集合:1298 篇

测试文本集合:584 篇

测试方法:三个类之间交叉很大,采用两种方法进行测试 1)采用自动分类 2)规则分类

测试结果:见表 1 和表 2,基于规则的分类的结果优于基于统计的分类的结果

表 1　基于统计的自动分类结果

	股市	利率	汇市
准确率	94.29%	83.57%	85.86%
召回率	96.77%	86.67%	78.70%

表 2　基于规则的自动分类结果

	股市	利率	汇市
准确率	97.41	93.57%	90.67%
召回率	99.41	97.04%	99.07%

测试 2:

数据来源:人民日报

类别:大学生　妇女问题　青少年犯罪　吸毒贩毒　失业与就业　环境污染　腐败问题

训练文本集合:1746 篇

测试文本集合:764 篇

测试方法:采用两种方法进行测试,1)采用基于统计的自动分类 2)采用基于规则的自动分类与基于统计的自动分类同时分,结果取二者的交集,来提高准确率。

测试结果:见表 3 和表 4,8 个类中交叉较大的类:妇女问题,失业与就业与大学生;青少年

维普资讯 http://www.cqvip.com

犯罪与吸毒贩毒;环境污染和腐败问题。采用方法2),明显提高了准确率。

表3 自动分类的结果

	住宅问题	吸毒贩毒	大学生	失业与就业	妇女问题	环境污染	腐败问题	青少年犯罪
准确率	96.12%	97.62%	94.17%	82.68%	92.16%	96.24%	94.87%	97.73%
召回率	99.00%	100%	97.98%	95.45%	94.00%	95.52%	92.50%	87.76%

表4 基于统计的自动分类与基于规则的自动分类交集的结果

	住宅问题	吸毒贩毒	大学生	失业与就业	妇女问题	环境污染	腐败问题	青少年犯罪
准确率	97.87%	100%	97.84%	99.00%	99.20%	100%	97.36%	100%
召回率	94.00%	97.00%	91.91%	91.91%	92%	80.02%	92.50%	85.71%

5 结语

本文介绍基于统计的自动分类的优势与不足,提出基于规则的分类对基于统计的分类的补充与改进的方法。设计实现了多分类器引擎的分类系统,性能良好。但是目前的分类器主要还是基于统计的思想,如何在汉语语言知识工程的基础上,真正实现基于自然语言理解的中文文本分类系统,高效地智能地实现中文文本分类,还需要中文信息处理的研究者们不断的努力!

参 考 文 献:

[1] 黄萱青,吴立德,石崎洋之,徐国伟.独立于语种的文本分类方法[J].中文信息学报.2000,14(6):1-7.

[2] Ji He, Ah-Hwee Tan, Chew-Lim Tan. A Comparative Study on Chinese Text Categorization Methods[J]. PRICAI Workshop on Text and Web Mining. 2000,24-35.

[3] 岳喜才,吴晓宇,郑崇勋,叶大田.一种大类别数分类的神经网络方法[J].计算机研究与发展.2000(3):278-283.

[4] 孙学刚,陈群秀,马亮.基于主题的Web文档聚类研究[J].中文信息学报.2003,17(3):21-26.

[5] 边肇祺,张学工.模式识别[M].第二版.北京:清华大学出版社.2000,284-304.

[6] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Feature. Proceedings of ECML-98,10th European Conference on Machine Learning[A]. In: Proceedings of ECML-98, 10th European Conference on Machine Learning[C]. Claire N line Rouveirol, 2000:137-142.

[7] 李辉,史忠植,许卓群.运用文本领域的常识改善基于支撑向量机的文本分类器性能[J].中文信息学报.2003,16(2):7-13.

[8] 王伟强,高文.段立娟. Internet上的文本数据挖掘[J].计算机科学.2000,14(4):32-36.

[9] 刁倩,王永成,张惠惠,何骥.文本自动分类中的词权重与分类算法[J].中文信息学报.2000,14(3):25-29.