基于SVM的中文文本分类技术研究

王金德 20721156

2008-04-29

摘 要

文本分类是文本挖掘的一个重要组成部分。通过文本分类，可以有效提高信息检索的速度和准确率。支持向量机（SVM）方法建立在统计学习理论（SLT）的VC维理论和结构风险最小化（SRM）原理之上，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。SVM针对于大样本集合，它把分类和降维结合在一起，非常适用于文本分类。

对于基于SVM的中文文本分类技术来说，它先对给定的中文文本进行分词、特征提取等操作，然后把文本用特征向量空间模型表示。这些文本特征向量就可以直接提交给SVM进行训练和分类。对于用来训练的文本特征向量，需要预先人工判断其类别，并一起提交给SVM。训练好的SVM就可以用来对实际的文本进行分类。

本文对基于SVM的中文文本分类技术进行了系统的研究，对该技术的每一步骤进行细化探讨，对它的特点进行了总结，为以后的相关研究提供一定的依据。

关键词：文本分类，分词，特征向量，SVM

目 录

第 1 章 概述

文本分类是在给定分类体系下，根据文本内容自动确定文本类别的过程[1]。由于人工分类非常费时，效率过低，90年代以来，众多的统计方法和机器学习方法应用于自动文本分类。

自动文本分类过程一般如下：首先对文本进行预处理，将文本用模型表示，进行特征提取；然后构造并训练分类器；最后用分类器对新文本进行分类[1]。

统计学习理论（SLT）是一种专门研究小样本情况下机器学习规律的理论。该理论针对小样本统计问题建立了一套新的理论体系，在这种体系下的统计推理规则不仅考虑了对渐近性能的要求，而且追求在有限信息的条件下得到最优结果[2]。

SVM建立在SLT的VC维理论和结构风险最小化（SRM）原理基础上，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力[2]。

SVM的主要优点在于：

- 专门针对有限样本情况，其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值。
- 算法最终将转化为一个二次型寻优问题，从理论上说，得到的将是全局最优点，解决了在神经网络方法中无法避免的局部极值问题。
- 算法将实际问题通过非线性变换转换到高维的特征空间，在高维特征空间中构造线性判别函数来实现原空间中的非线性判别函数，从而保证学习机器有较好的推广能力，同时它巧妙地解决了维数问题，其算法复杂度与样本维数无关。

根据SVM的特点，文本实质上是一种高维数据，所以SVM很适合于文本分类[3,4]。基于SVM的中文文本分类器首行对文本进行分词，再利用如信息论中的互信息量等概念抽取特征，把文本表示成特征向量空间模型，计算文本向量中每个特征的权重，并利用SVM作为分类器进行文本分类。

第 2 章　基于SVM的中文文本分类技术

2.1　中文文本分类技术概述

目前典型的中文文本分类技术过程如图2.1:

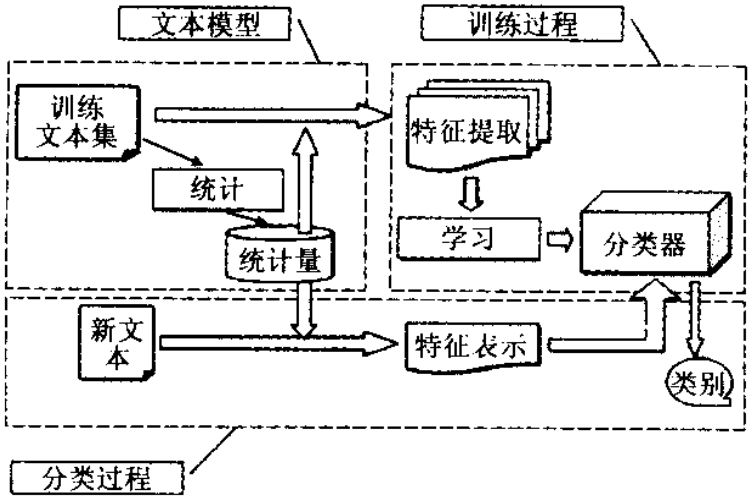

图 2.1　自动文本分类过程[1]

该技术详细步骤如下：

1. 准备测试集文本；
2. 测试集文本人工分成若干类别；
3. 利用词典进行文本分词；
4. 去停用词、合并数字人名，进行一定的降维；
5. 采用TF-IDF公式计算词条权重；
6. 特征项抽取（利用互信息量、词熵、X2 统计量等方法）；
7. 特征降维处理，方法有主成分分析、潜在语义标引、非负矩阵分解）；
8. 构造分类器（比如SVM、向量距离分类法、KNN算法、贝叶斯算法等）；
9. 用特征词汇描述测试文本；
10. 取部分测试文本用来训练分类器，另一部分测试，并根据结果进行一些改进和优化；
11. 利用训练好的分类器对新文本进行分类。

在该技术中，所使用的关键技术主要包括倒排索引建立（包括中文分词）、特征向量提取和SVM中文文本分类。

2.2　关键技术分析

2.2.1　倒排索引建立

倒排索引文件的建立，为特征向量提取提供必要的数据结构。在建立倒排索引文件之前，需要对中文文本进行分词。

所谓的中文分词是一种将连续的汉语文本序列按一定规则拆分为具有独立语义的词组的过程。[1] 中文分词是当前分词技术中的一种，分词技术从语言文本结构上来讲大致有两类：一类以英文为代表的西方语言文本，其文本中的词组以空格作为自然间隔，从语义准确性及技术复杂度来讲都比较简单。另一类是以汉语为代表的东亚语言文本，由于文本是由连续文字组成，缺乏有效的间隔，虽有句、段分隔，但在进行机器语言学习、文本语义理解分析过程中都需以词组为最小单位。因此东亚文本语言实现分词技术相对西方文本语言来讲，更加的复杂和困难。

在实际的中文分词中，可以根据一些分隔词（Stop Words），对文本进行分词，得到一个中文词语的列表。在得到分词结果之后，要对分词的结果列表进行处理，主要是去掉文本中对分类无用的词语，如叹词，助词等，去掉停用词（stop word）和标点符号，合并数字人名等，同时获得保留下来的词语的词频。经过这些处理之后的词语列表，就为创建倒排文件做好了准备。

在文本分类中创建倒排索引文件是为了提高提取特征向量和计算特征向量权重的效率。因为在倒排索引文件中可以在O(1)时间复杂度下提供对特征向量选择，特征权重计算所需要的重要变量值，如每一个关键词的总词频和文件词频，出现某个特征向量的文件的个数等。

2.2.2　特征向量提取

特征向量提取过程主要是把中文文本转化成用特征向量权重表示的文件，便于SVM对文本进行分类。而特征向量的维在文本分类中是指在文本中出现，可以标识文本所属类别的词语。特征向量提取过程可以分为特征抽取和生成文

本特征向量。

特征的抽取一般是通过构造一个特征评分函数，把测量空间的数据（分词词语）投影到特征空间，得到在特征空间中的值，然后根据这个值对每个特征进行评估，特征选择就成了选择值最高的若干个特征[5,6]。

选择合适的特征评分函数也是文本分类的一个关键内容。在中文文本分类中，常用的特征评分函数主要有如下几种[7]：

- 词条和类别的互信息。词条和类别的互信息体现了词条和类别的相关程度，互信息越大，词条和类别的相关程度也越大。
- 词条的X统计。词条的X统计比较了词条对一个类别的贡献和对其余类别的贡献的大小，以及词条和其余词条对分类的影响。
- 词条的期望交叉熵。交叉熵反映了文本类别的概率分布和在出现了某个特定词的条件下文本类别的概率分布之间的距离，词条的交叉熵越大，对文本类别分布的影响也越大。
- 文本证据权。它比较了类出现的概率和在给定特征下类出现的条件概率之间的差别。
- 信息增益。它和期望交叉熵不同之处仅在于它还考虑了词条未出现的情况。

特征抽取成功之后，把文本表示成文本特征向量一般分为两步：

1. 计算每个词条在各个类别中的互信息量。
2. 依次抽取每个类别中互信息量最大并且不在特征集合中的词，加入到特征集合中。

这两个步骤不断迭代抽取，直到抽取的特征个数等于用户配置的最大特征个数或者抽取了所有特征为止。

事实上，把文本表示成文本特征向量就是把中文文本转化成可以被SVM分类器识别的特征权重表示。对于文本来说，计算权重常采用TF*IDF技术实现。这样，一般地生成过程为：

首先将要分类的文本进行分词，然后判断每一个词语是否属于特征向量，如果这个词语属于特征向量的一个维度，就计算的它的TF*IDF的值作为它的特征向量权重。整个过程如图2.2:

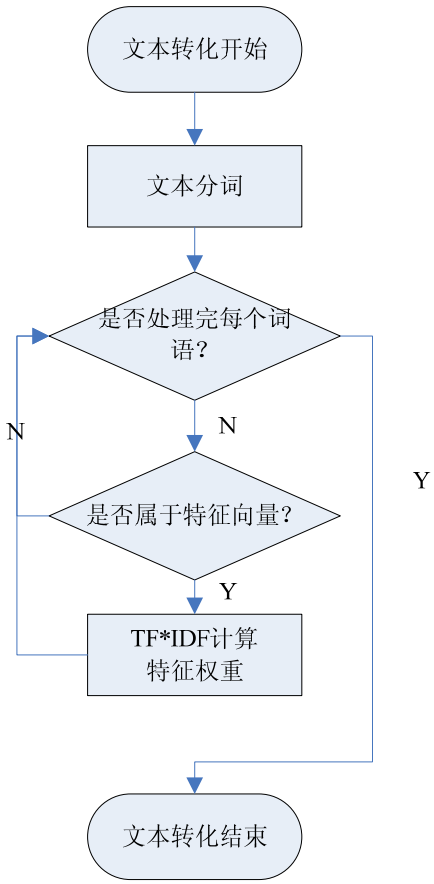

图 2.2　文本转化过程

2.2.3　基于SVM中文文本分类

把文本表示成文本特征向量之后，就可以使用SVM进行训练和实际分类了。文本特征向量代表了一个文本，同时能被SVM所识别，它相当于实际文本和SVM之间的媒介。具体的分类过程如图2.3：

2.3　技术评价

基于SVM的中文文本分类技术，与其它同样功能的分类技术相比，表现出了优秀的推广性能。并且SVM正在不断地发展中[8,9]，包括块算法、固定工作样本集、序列优化思想等不断融入到SVM中，这也促进了基于SVM的中文文本分类技术的发展。

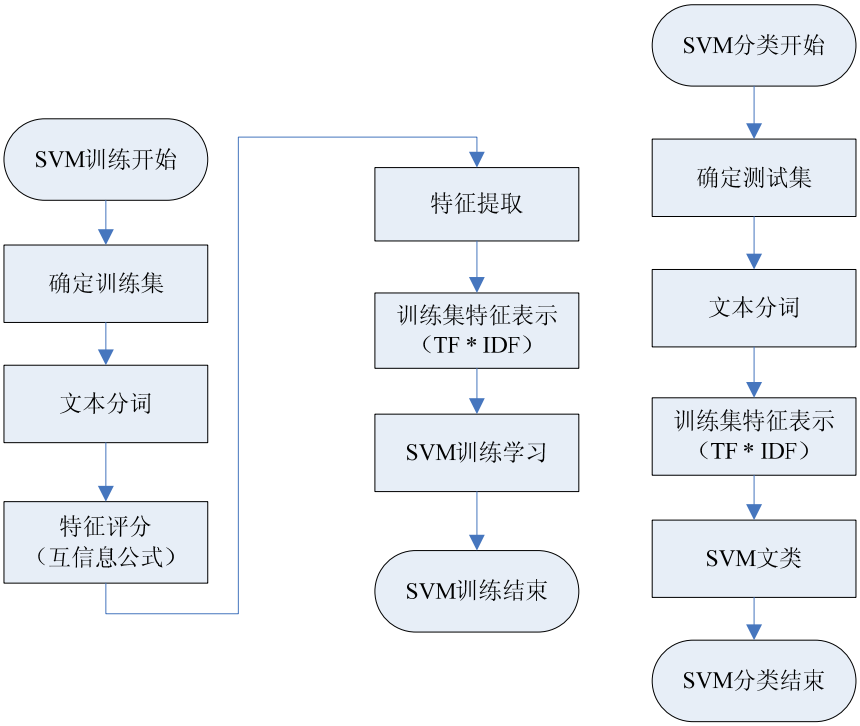

图 2.3　基于SVM中文文本分类器分类过程

不过该技术在计算上存在着一些问题，包括训练算法速度慢、算法复杂而难以实现、及分类运算量大等。使用SVM进行中文文本分类，速度慢的原因主要在于：

- SVM方法需要计算和存储核函数矩阵，当样本点数较大时，需要很大的内存。
- SVM在二次型寻优过程中要进行大量的矩阵运算，多数情况下，寻优算法是占用算法时间的主要部分。

另外，中文分词的技术效果，也是中文文本分类的关键。

第 3 章　总结

文本分类是在给定分类体系下，根据文本内容自动确定文本类别的过程。SVM建立在统计学习理论的VC维理论和结构风险最小化（SRM）原理基础上，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。SVM适用于高维数据，而文本正是一种高维数据，于是，基于SVM的中文文本分类表现出优良的特性。

本文深入研究了基于SVM的中文文本分类技术，对该技术的每一步骤进行细化探讨。该技术的过程主要为：首先对文本进行预处理，将文本用模型表示，进行特征提取；然后构造并训练SVM分类器；最后用SVM分类器对新文本进行分类。

本文对该技术的缺点进行了分析，得出了一些结果。中文分词并不完善，它本身是一个正不断研究中的课题。SVM技术也存在着一些问题，包括训练算法速度慢、算法复杂而难以实现、及分类运算量大等。而一些新的研究把块算法、固定工作样本集、序列优化思想等不断融入到SVM中，使SVM克服一些缺陷，表现出更优良的特性。这些基础技术的提高将有力地推动基于SVM的中文文本分类技术的发展。

参考文献

[1] 高洁. 文本分类技术研究. 计算机应用研究, 2004

[2] Burges C J C. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., 1998, 2(2):121–167

[3] Joachims T. A statistical learning learning model of text classification for support vector machines. In: Proceedings of SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA: ACM, 2001. 128–136

[4] Tong S and Koller D. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2002, 2:45–66

[5] Lewis D D. Feature selection and feature extraction for text categorization. In: Proceedings of HLT '91: Proceedings of the workshop on Speech and Natural Language, Morristown, NJ, USA: Association for Computational Linguistics, 1992. 212–217

[6] Yang Y and Pedersen J O. A comparative study on feature selection in text categorization. In: Proceedings of ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. 412–420

[7] 文本分类中的特征抽取. 秦进, 陈笑蓉, 汪维家, 陆汝占. 计算机应用, 2003, 23

[8] Liu Y, You Z, and Cao L. A novel and quick svm-based multi-class classifier. Pattern Recogn., 2006, 39(11):2258–2264

[9] wen Chen X, Zeng X, and van Alphen D. Multi-class feature selection for texture classification. Pattern Recogn. Lett., 2006, 27(14):1685–1691