基于 SVM 的分类方法综述

张小艳 李 强
(西安科技大学计算机系 陕西 西安 710054)

【摘 要】本文介绍了文本分类的起源,常用的几类文本分类方法及基于 SVM(Support Vector Machines)文本分类的基本原理和方法。并在分析文本分类的特点的基础上比较了在文本分类中应用 SVM 的优势及存在的问题。最后总结出了 SVM 在文本分类中应用的两个主要研究方向。

【关键词】支持向量机;文本分类;机器学习

The Summary of Text Classification Based on Support Vector Machines

ZHANG Xiaoyan LI Qiang
(Dept. of computer, Xi'an University of Science & Technology. Xi'an 710054,China)

【Abstract】In this paper, the origin of text classification and the commonly used types of text classification are briefly introduced. The principles and methods of text classification based on SVM(Support Vector Machines) are introduced in detail. The advantages and disadvantages of SVM used in text classification are also explored. Finally, main research directions of text classification based on SVM are discussed.

【Key words】support vector machines; text categorization; Machine Learning

1.引言

自动文本分类的研究最早可以追溯到二十世纪六十年代 Maron 的研究工作。到二十世纪八十年代之前,在自动文本分类方面占主导地位的一直是基于知识工程的分类方法。基于知识工程的方法存在分类规则制定困难、推广性差的缺点,因此很难大规模推广应用。二十世纪九十年代以来,随着信息存储技术和通信技术的迅猛发展,大量的文字信息开始以计算机可读的形式存在,并且其数量每天仍在急剧增加。这一方面增加了对于快速、自动文本分类的迫切需求,另一方面又为基于机器学习的文本分类方法准备了充分的资源。在这种情况下,以机器学习技术为主的信息分类技术逐渐取代了基于知识工程的方法,成为自动文本分类的主流技术[1]。

常用的自动文本分类算法主要包括三大类。一类是基于概率和信息理论的分类算法,如朴素贝叶斯算法(Naive Bayes, 简称 NB)[2],最大熵算法(MaximumEntropy)等[3];另一类是基于 TFIDF 权值计算方法的分类算法,这类算法包括 Rocchio 算法,TFIDF 算法,k 近邻算法(k Nearest Neighbors, 简称 kNN)等[4];第三类是基于知识学习的分类算法,如决策树(Decision Tree),人工神经网络(Artificial Neural Networks, 简称 ANN),支持向量机(Support Vector Machine,简称 SVM)等算法[5]。本文主要介绍基于 SVM 的分类方法。

2.支持向量机

2.1 支持向量机简介

SVM 是在高维特征空间使用线性函数假设空间的学习系统,它集成了最大间隔超平面、Mercer 核、凸二次规划、稀疏矩阵核松弛变量等多项技术[6]。训练集是训练文本的集合,通常表示为:

$S=((x_i,y_i),L(x_i,y_i))\subset(X\times Y)^l$

其中 l 是文本数目,x_i 指文本,y_i 是它们的标记,X 表示输入空间,Y 表示输出域。

如图 1 所示,假设空心和实心点表示两类训练样本,实线为分类面,虚线为平行于实线的平面,并且是经过两类训练样本中离分类面最近的平面。如果训练集中的所有训练样本均能被某超平面正确划分,且距该平面最近的异类向量之间的距离最大,则该平面为最优超平面[7],用 $\langle\omega\cdot x+b\rangle=0$ 表示。

其中 ω为分类面的法线,b 为分类面的偏移量, 向量 x 位于分类面上,其中位于虚线上的向量被称为支持向量(图 1 中加圈的点)。

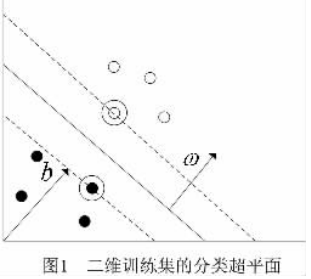

图1 二维训练集的分类超平面

2.2 分类原理

SVM 的分类原理就是在线性可分的情况下寻找一个最优超平面,使其在误判率最低的前提下达到最优的分类效果.假设给定一个线性可分的训练文本:

$S=((x_i,y_i),L(x_i,y_i))\subset(X\times Y)^l$

求解优化问题:$\min imise_{\omega,b}\ \langle\omega\cdot\omega\rangle$

subject to $y_i(\langle\omega\cdot x\rangle+b)\geq 1, i=1,L,l$

可以得到超平面$(\omega\cdot b)$,它实现了几何间隔 $\gamma=||\omega||_2$ 的最大间隔超平面。

对于该问题的求解是典型的有约束二次规划问题,采用拉格朗日乘数法可转化为对偶问题[6]。其相应的对偶形式:

$$\omega=\sum_{i=1}^{l}\alpha_i y_i x_i,(\alpha_i\geq 0),i=1,K,l$$

$$0=\sum_{i=1}^{l}\alpha_i y_i x_i,(\alpha_i\geq 0),i=1,K,l$$

代入到原始的拉格朗日函数,得到:

$$L(\omega,b,\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\alpha_i\alpha_j y_i y_j\langle x_i\cdot x_j\rangle$$

假定参数 α^* 和 b^* 是下面二次优化问题解:

$$\max imize\ W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\alpha_i\alpha_j y_i y_j\langle x_i\cdot x_j\rangle$$

$$\text{subject to } 0=\sum_{i=1}^{l}\alpha_i y_i,(\alpha_i\geq 0),i=1,K,l$$

则决策规划由 sgn(f(x))给出,这里 $f(x)=\sum_{i=1}^{l}\alpha^* y_i\langle x_i\cdot x_j\rangle+b^*$

若判定一个文本所属的类别,可把 x 代入 sgn(f(x))中.当 sgn(f(x))=1,则 x 属于该类;否则不属于该类。

对于非线性问题,需要引入松弛变量,它允许在一定程度上违反间隔约束[6]。

$\min imise_{\omega,b}\ \langle\omega\cdot\omega\rangle$

则原始的优化问题变为如下:

subject to $y_i(\langle\omega\cdot x\rangle+b)\geq 1-\xi_i, i=1,L,l$

$\xi_i\geq 0, i=1,L,l$

可以通过选择使用恰当的核函数来替换内积,隐式地将训练文本非线性映射到高维空间,使其在高维空间中线性可分,而不增加可调参数的个数。设线性映射 $\Phi:X\to H$ 将输入空间的文本映射到高维特征空间 H 中。当在 H 中构造最优超平面时,训练算法仅使用空间点积,即 $\Phi(x_i)=\Phi(y_i)$[7],则核函数 $k(x_i,y_i)=\Phi(x_i)\cdot\Phi(y_i)$.此时决策规则 sgn(f(x))中的 f(x)为:

$$f(x)=\sum_{i=1}^{l}\alpha^* y_i\langle x_i\cdot x\rangle+b^*$$

3.基于 SVM 文本分类方法的优势

文本分类的特点:①文本分类所需要处理的样本空间非常庞大,

即便通过简化，仅考虑词，而不考虑词的次序不同、断句等的不同所表达的含义的不同，样本的维数也很高；②文本向量非常稀疏；③文本特征之间存在较大的相关性；④文本训练样本集中存在较多噪音；⑤不同文本类别的训练样本数目往往存在较大差别。文本分类的这些特点，使得很多分类算法的效果不好。与其它文本分类算法相比，支持向量机主要具有如下优势：

1)文本数据向量维数很高，对于高维问题，支持向量机具有其它机器学习方法不可比拟的优势；

2)文本向量特征相关性大，许多文本分类算法建立在特征独立性假设基础上，受特征相关性的影响较大，而支持向量机对于特征相关性不敏感；

3)文本向量存在高维稀疏问题，一些文本分类算法不同时适合于稠密特征矢量与稀疏特征矢量的情况，但支持向量机对此不敏感；

4)文本分类样本收集困难、内容变化迅速，支持向量机能够找出包含重要分类信息的支持向量，是强有力的增量学习和主动学习工具，在文本分类中具有很大的应用潜力。

4.基于 SVM 文本分类方法中存在的问题

支持向量机从被广泛重视到现在只有几年的时间，已经提出的很多关于支持向量机的训练算法，从训练时间和分类精度两种角度进行优化。其中还存在很多尚未解决或尚未充分解决的问题，需要进一步完善和改进以适应实际应用的需要，支持向量机在实际应用中、尤其在文本分类等类别和样本数目多、噪音多的应用中存在的主要问题包括：

1)对于需要求解二次规划问题的支持向量机模型，当样本数目较多时，其训练速度较慢。尤其对于训练样本和支持向量数目多的分类问题，支持向量机的分类速度过慢，这一点限制了支持向量机的应用，成为支持向量机方法进入大规模实用化阶段的瓶颈。如何进一步改进和完善支持向量机模型及其训练算法，是支持向量机研究中的热点问题。

2)支持向量机是针对两类分类问题提出的，用于多类分类必须将其推广。对于类别数目较多的分类问题，目前仍缺乏有效的支持向量机多类分类方法。

3)支持向量机中核函数及参数的选择没有好的确定的方法，仍然凭经验寻求。不同的核函数对应的支持向量集合有所不同，而支持向量的少量丢失都会引起分类精度的下降，大规模文本分类中在减少样本数目时怎样保证不丢失支持向量仍然是个难点。

5.总结

总体来说，目前，支持向量机在文本分类中的应用研究主要包括两方面的内容：一是利用支持向量机的优势，挖掘支持向量机在文本分类中的应用潜力，解决文本分类中存在的问题。例如，支持向量机能够找出包含重要分类信息的支持向量，是强有力的增量学习和主动学习的工具等；二是研究支持向量在文本分类应用中存在的尚未解决或尚未完全解决的问题，针对文本分类的特点，提出提高支持向量机在文本分类中的应用效果的新方法。科

参考文献】

[1] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报,2006, 17(9):1848-1859.
[2] McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. Learning for Text Categorization: Papers from the AAAI Workshop. Tech.rep. WS-98-05, AAAI Press, 1998:41-48.
[3] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: IJCAI99 Workshop on Machine Learning for Information Filtering. 1999, 61-67.
[4] Yang Y, Chut C. An Example-Based Mapping Method for Text Classification and Retrieval. ACM Transactions on Information Systems, 1994, 23(3): 252-277.
[5] 邓乃扬, 田英杰.数据挖掘中的新方法—支持向量机. 第一版. 北京: 科学出版社, 2004, 1-223.
[6] Nello Cristianini, John Shawe-Taylo. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M].北京:机械工业出版社, 2005.
[7] 瞿林, 刘亚军. 支持向量机的中文文本分类研究 [J]. 计算机与数字工程. 2005,33(3):21-23.

作者简介：张小艳（1967—），女，陕西西安人，硕士，副教授。主要研究领域：网络集成与数据库技术、知识工程与智能系统等。
李强（1982—），男，硕士研究生。主要研究领域：网络集成与数据库技术、知识工程与智能系统。

※基金项目：陕西省教育厅专项课题（06JK248）资助。

[责任编辑：张新雷]

（上接第 346 页）教师，有针对性地开设一些实践技能培养为重点的训练。培训方式可采用校内训练和校外训练两种。校内培训是利用寒暑假聘请专家或生产第一线的具有丰富实践经验的专业技术人员，来校培训；校外培训采用有计划选派中青年教师到生产第一线的养殖企业挂职培训。提高专业教师实际生产实践能力和对新技术、新工艺、新方法的应用能力。

2. 积极鼓励和支持专业教师参加生产实践　学校要制定详细的计划，有针对性的要求青年和中年教师完成生产实践锻炼的具体内容，分析和掌握生产实践中的第一手资料和信息，并写出书面汇报材料。

随着现代养殖业的发展，养殖业生产已发生了根本的变化，大规模、集约化、信息化的生产是行业的主要生产形式，作为从业的生产第一线技术人员，为适应工作要求，更要全面掌握规模化、集约化生产和管理技术，掌握对畜禽群发疾病、流行病的预防和治疗技术等知识，而不是以教科书为主进行专业教学。因此，从专业知识的角度发生了重点的转移。这就要求畜牧兽医专业教师在日常教学工作中既要讲授基本理论；又要手把手的教学生岗前职业训练，保证毕业生具有独立从事畜牧业生产岗位工作的能力。

3.走“工学”结合的路子　“工学”结合是对专业教师进行继续教育、提高专业素质和专业能力的有效途径。农业高职院校要加强与生产、企业及科研院所的联系，建立“工学”基地，定期组织专业教师到基地学习、生活、实践，接受新的专业知识和信息，全面掌握专业领域科学技术发展趋向，了解生产一线和社会对本专业的需求，从而收集教学素材，丰富教学内容。通过“校企合作、‘工学’结合的方式积极开展技术攻关或产品研发。

4.强化组织管理　由于专业教师深入生产第一线亲自参加生产、摸爬滚打是一种艰苦的生活体验，所以靠专业教师自发的或积极主动的奔赴生产第一线参加锻炼的人，可以肯定的说几乎没有。因此，学院要设立和制订相应的组织管理机构、管理措施及奖励政策，如提高补贴、补助费等；另外学院要组织有关部门经常赴生产第一线看望关心和勉励他们，对在生产实践期间表现出色的专业教师要予以表彰通报，对在实践锻炼期间发表论文（一般期刊或核心期刊）要给予重奖，并且作为评优、评先进、评职称、晋升的量化指标。科

参考文献】

[1] 彭宝利.付云强.雷晓忠等.高职院校教师应深入企业调研[J].中国职业技术教育,2005(11):35-36.
[2] 宋继东.宋晓燕.我国高职院校存在的突出问题及对策[J]中国成人教育 2008(12):91-9.
[3] 郭朝辉.提高“双师型”教师比例的对策[J]职业技术教育,2007(10):29-30.
[4] 刘振湘.何华西.高职高专畜牧兽医专业人才培养方案研究[J]高等农业教育,2003(12):79-84.
[5] 孙玲.高职院校师资队伍建设方案研究[J].教育与职业,2007(33):143-144.

作者简介：赵政（1957—），男，山东潍坊人，广西农业职业技术学院，副教授，从事畜牧兽医专业教学和科研工作。

※基金项目：广西农业职业技术学院教育科学研究课题（编号 YJJ0802）。

[责任编辑：张艳芳]