多类支持向量机文本分类方法

张 苗, 张德贤
(河南工业大学 信息科学与工程学院, 河南 郑州 450051)

摘 要: 文本分类是数据挖掘的基础和核心, 支持向量机(SVM) 是解决文本分类问题的最好算法之一。传统的支持向量机是两类分类问题, 如何有效地将其推广到多类分类问题仍是一项有待研究的课题。介绍了支持向量机的基本原理, 对现有主要的多类支持向量机文本分类算法进行了讨论和比较。提出了多类支持向量机文本分类中存在的问题和今后的发展。

关键词: 文本分类; 机器学习; 支持向量机; 多类支持向量机
中图分类号: TP311 文献标识码: A 文章编号: 1673- 629X(2008)03- 0139- 03

Research on Text Categorization Based on M- SVMs

ZHANG Miao, ZHANG De-xian
(College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450051, China)

Abstract: Text categorization is the basis and core of textual data mining and SVM is one of the best solutions to this question. How to effectively extend support vector machines (SVM) for multi- category classification is still an on- going research issue. Presents a general overview of existing representative methods for multi- category support vector machines and systematically compares them. Lastly, it presents some existing problems and future developments in text categorization field.

Key words: text categorization; machine learning; SVM; multi- category SVM

0 引 言

自动文本分类(Text Categorization, TC) 是一个有监督的学习任务, 它定义为: 根据一些已经分配好类标签(这些类标签预先定义好) 的训练文档集合, 来对新文档分配类标签。基于机器学习的文本分类系统能够在给定的分类模型下, 根据文本的内容自动对文本分门别类, 从而更好地帮助人们组织文本、挖掘文本信息, 因此得到日益广泛的关注, 成为信息处理领域最重要的研究方向之一。

文本分类过程中, 经过预处理(如分词、去停词和去标点等过程) 的学习样本进行特征选择(feature selection) 后, 每个文本被表示为一个当前 n 维特征向量空间中的向量, 作为机器学习算法的输入。用于文本分类的机器学习方法有 kNN、Bayes、Rocchio、神经网络以及支持向量机(Support Vector Machines, SVM) 等。其中, SVM 反映了当前文本分类方法的性能水平[1]。

收稿日期: 2007- 06- 07
基金项目: 河南省科技攻关项目(0324220024)
作者简介: 张 苗(1979-), 女, 河南尉氏人, 讲师, 硕士研究生, 研究方向为模式识别; 张德贤, 教授, 博士, 研究方向为计算机智能技术。

1 SVM 基本原理

SVM 是 Vapnik 等人根据统计学习理论提出的一种新的机器学习方法[2], 它以结构风险最小化原则为理论基础, 通过适当选择函数子集及该子集中的判别函数使学习机的实际风险达到最小, 保证了通过有限训练样本得到的小误差分类器对独立测试集的测试误差仍然小, 得到一个具有最优分类能力和推广泛化能力的学习机。SVM 的基本思想是, 对于一个给定的具有有限数量训练样本的学习任务, 任何对于给定训练集和机器无错误地学习任意训练集的能力进行折中, 以得到最佳的性能。SVM 在模式识别中的思想是构造一个超平面作为决策平面, 使正负之间的空白最大。模式识别的主要任务是构造一个目标函数, 使两类模式正确地分开[3]。在线性可分和线性不可分的情况下, 可转换为一个典型的二次规划问题。在非线性时, 利用 Mercer 核非线性映射将输入矢量映射到一个高维的特征空间, 在高维的特征空间上构造最佳的超平面, 然后采用线性分类器分类。

2 多类支持向量机文本分类

V. Vapnik 提出的支持向量机理论因其坚实的理

论基础和诸多良好特性在近年获得了广泛的关注。支持向量机最初是为两类分类问题而设计的,而在实际应用中,多类分类问题更为普遍,文本分类就是一个多类分类问题。因此,如何在将支持向量机的优良性能推广到文本分类中的同时提高支持向量机的训练和决策速度及解决目前支持向量机多类分类方法中不可分区域的分类问题便成为一项有实际意义的研究课题。

当前已经有许多算法将 SVM 推广到多类分类问题,这些算法统称为"多类支持向量机"(Multi- Category Support Vector Machines, M- SVMs)[4]。目前常用于文本分类的主要有以下几类[5,6]。

2.1 组合法

2.1.1 1- a- r(一对多)方法

SVM 多类分类方法最早使用的算法就是"一对多"方法[7]。要得到多类分类器,通常的方法是构造一系列两类分类器,其中的每一个分类器都把其中的一类同余下的各类分划开。然后据此推断输入 X 的归属。"一对多"方法是对于 k 类问题构造 k 个SVM 子分类器。在构造第 i 个 SVM 子分类器时,将属于第 i 类别的样本数据标记为正类,不属于 i 类别的样本数据标记为负类。测试时,对测试数据分别计算各个子分类器的决策函数值,并选取函数值最大所对应的类别为测试数据的类别。第 i 个 SVM 需要解决下面的最优化问题:

$$\text{Minimize: } \frac{1}{2}(w^i)^{\mathrm{T}}w^i + C\sum_{j=1}^{l}\xi_j^i$$

$$\text{Subject to: } (w^i)^{\mathrm{T}}\varphi(x_j) + b^i \geqslant 1 - \xi_j^i$$

$$\text{if } y_j = i,\ (w^i)^{\mathrm{T}}\varphi(x_j) + b^i \leqslant 1 + \xi_j^i$$

$$\text{if } y_j \neq i,\ \xi_j^i \geqslant 0,\ j = 1, 2, \cdots, l$$

解决以上最优化问题后,就可以得到 k 个决策函数:

$$(w^i)^{\mathrm{T}}\varphi(x) + b^i,\ i = 1, 2, \cdots, k$$

对于待测样本 x,将其输入这 k 个决策函数中,得到 k 个值,取得最大值的函数对应的类别即为该样本所属类别。

1- a- r 方法因简单易于实现而得到了广泛的应用,但是,它也存在许多缺点,这一分类算法泛化能力较差,并且训练时间与训练样本类别数 k 成正比,当训练样本数目大时,训练困难。尤其是一对多方法会造成训练集不均匀(如果类别之间不均匀,在每个两类分类器中负类别的样本将大大多于正类别的样本),对小样本的类别识别精度比较低如图1所示,区域 A, B, C 和 D 为 1- a- r 方法的不可识别域。

2.1.2 1- a- 1(一对一)方法

这种方法也是基于两类问题的分类方法,不过这里的两类问题是从原来的多类问题中抽取的。具体做法是:"一对一"方法(One- against- one Method)是分别选取 2 个不同类别构成一个 SVM 子分类器,这样共有 $k(k-1)/2$ 个 SVM 子分类器[8]。在构造类别 i 和类别 j 的 SVM 子分类器时,在样本数据集选取属于类别 i、类别 j 的样本数据作为训练样本数据,并将属于类别 i 的数据标记为正,将属于类别 j 的数据标记为负。

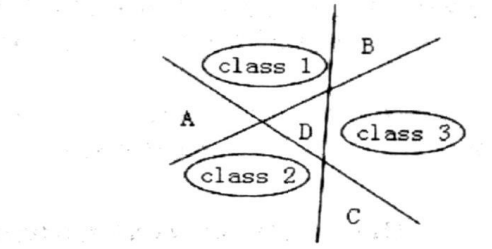

图 1 1- a- r 分类示意图

"一对一"方法需要解决如下的最优化问题:

$$\text{Minimize: } \frac{1}{2}(w^{ij})^{\mathrm{T}}w^{ij} + C\sum_t \xi^{ij}$$

$$\text{Subject to: } (w^{ij})^{\mathrm{T}}\varphi(x_t) + b^{ij} \leqslant 1 - \xi^{ij},$$

$$\text{if } y_t = i,\ (w^{ij})^{\mathrm{T}}\varphi(x_t) + b^{ij} \leqslant 1 + \xi^{ij}$$

$$\text{if } y_t = j,\ \xi^{ij} \geqslant 0$$

解决这一最优化问题后,也即用训练样本进行训练后就可以得到 $k(k-1)/2$ 个 SVM 子分类器。测试时,将测试数据对 $k(k-1)/2$ 个 SVM 子分类器分别进行测试,并累计各类别的得分,选择得分最高者所对应的类别为测试数据的类别。每一类与其他各类别分别构成一个两类问题,k 个类别共构造 $k(k-1)/2$ 个两类 SVM。与一对多模式类似,训练样本也需要相应地改变类别标签。预测时,样本经过所有的两类 SVM,得到 $k(k-1)/2$ 个识别结果;采用投票法来决定测试样本的类别,即在 $k(k-1)/2$ 个分类函数中,出现最多次数的类别就是最终的预测类别。

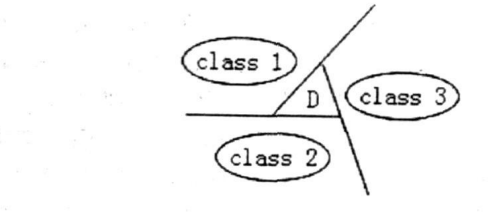

图 2 1- a- 1 分类示意图

1- a- 1 方法的分类示意图如图 2 所示,存在不可识别区域 D;需要构造的支持向量机数目较多,对于类别数目多的分类问题,训练速度较低,分类函数随着类别数的增加迅速增加,从而使预测过程变慢。但是文献中表明采用一对一算法往往具有较高的分类精度。

2.2 决策树分类法

支持向量机通常和决策树结合起来, 构成多类别的识别器。SVM 决策树具有层次结构, 每个层次子 SVM 的级别和重要性不相同, 其训练集合的构成也不同; 测试是按照层次完成的, 对某个输入样本, 可能使用的子 SVM 数目介于 1 和决策树的深度之间, 测试速度快; 决策树各节点和树叶的划分没有理论指导, 需一定的先验知识。这种方法同时处理所有的样本和类别, 没有忽略为得到每一问题最佳解决方案的任何相关信息[9]。除此之外, 所得的向量机需要的支持向量个数很少, 而且在训练集可分时能够达到很好的性能。不可分时, 错分的样本将被多次惩罚, 得出一个针对这些样本有偏支持向量机的研究进展的解决方案。

树型支持向量机多类分类方法的主要优点是需要训练的支持向量机数目和各支持向量机的训练样本数目都较少, 并且分类时也不必遍历所有的支持向量机分类器, 具有较高的训练速度和分类速度, 对于类别数目多的分类问题, 它的优势更为明显。但是这种方法如果在决策树的某个节点上发生了分类错误, 将会把错误延续下去, 该节点后续下一节点的分类就失去了意义。决策树分类 SVM 决策图如图 3 所示。

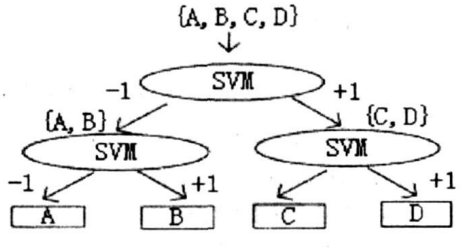

图 3　决策树分类支持向量机决策图

2.3 DDAG

决策导向非循环图(Decision Directed Acyclic Graphs, DDAG) 方法是将多个两类分类器组合成多类分类器。在训练阶段, 其与 1- a- 1 方法相同, 对于 K 类问题, DDAG 含有 $K(K-1)/2$ 个二类分类器。而在决策阶段, 使用从根节点开始的导向非循环图, 具有 $K(K-1)/2$ 个内部节点以及 K 个叶子节点, 每个内部节点都是一个二类分类器, 叶子节点为最终的类值。对一个测试样本, 从根节点开始根据分类器的输出值决定其走左侧或右侧路经, 一直到叶子节点得 到样本所属的类值为止。DDAG 多类 SVM 决策图如图 4 所示。

该方法的主要优点是采用了有向无环图的组合策略, 分类时不必遍历所有的分类器, 具有更高的分类效率, 决策速度比 1- a- r 方法或 1- a- 1 的投票方法快, 而在训练阶段其速度与 1- a- 1 相同, 因此总的来说, DDAG 的速度是这 3 种方法中最快的。主要缺点是需要构造的支持向量机数目较多, 对于类别数目多的分类问题, 训练速度较低。由于泛化能力较好, 多类支持向量机比其他方法识别效果更好, 在众多领域得到了广泛的应用, 并在应用的基础上得到了进一步的发展。

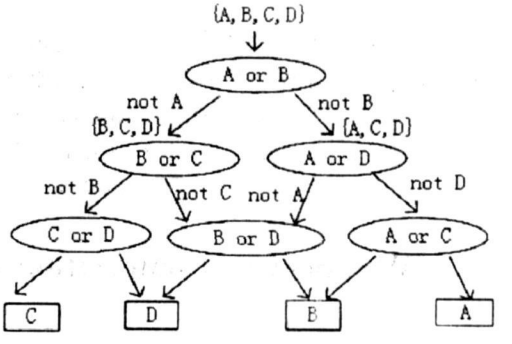

图 4　DDAG 多类支持向量机决策图

多类支持向量机文本分类算法比较如表 1 所示。

表 1　多类支持向量机分类算法比较

	SVM 个数	训练代价	学习代价	正确率
一对多	k	一般	较大	一般
一对一	$k(k-1)/2$	较小	一般	较高
决策树	$k-1$	较小	一般	一般
DDAG	$k(k-1)/2$	较大	较小	较高

3 总结和展望

文本分类是一个多分类问题, 随着支持向量机在此领域的深入运用, 如何有效地将该方法推广到多类分类问题中必将引起人们广泛的兴趣。该文讨论了支持向量机的基本原理及现有主要的支持向量机多类分类算法, 系统地比较了各算法的性能。从以上的分析看, 1- a- r, 1- a- 1, DDAG SVMs 三种支持向量机多分类方法存在不可分盲点, 需要训练的支持向量机个数太多, 有的分类未知样本时使用的支持向量机过多, 目前决策树支持向量机方法在文本分类中使用较为广泛, 并和聚类、模糊集、遗传算法等相结合, 在分类的准确性和时间效率方面均有所提高。对于多类支持向量机而言, 采用不同的拓扑结构将形成不同的算法, 这些算法的执行效率受模型拓扑结构的制约, 这也为进一步改进算法提供了有效途径。

参考文献:

[1] Debole F, Sebastiani F. An analysis of the relative hardness of reuters- 21578 subsets[J]. Journal of the American Society for Information Science and Technology, 2004, 56(6): 584- 596.

(下转第 156 页)

contained in c;// 对包含于 c 中 C_k 内的所有候选者记数

L_k = Candidates in C_k with minimum support;// L_k = C_k 中满足最小支持度的候选者

end

Answer = Maximal Sequences in $\bigcup_k L_k$;

上面算法的关键是候选集的产生, 具体候选者的产生如下:

AprioriALL- generate( ) 函数// 计算候选者的产生

输入: 所有的大(k - 1) 序列的集合 L_{k-1}

输出: 候选 C_k

insert into C_k// 首先进行 L_{k-1} 与 L_{k-1} 的连接运算

select p.litemset$_1$, p.litemset$_2$, …, p.litemset$_{k-1}$, q.litemset$_{k-1}$

from $L_{k-1}p$, $L_{k-1}q$//p, q 是L_{k-1} 中不同的序列串

where p.litemset$_1$ = q.litemset$_1$, …, p.litemset$_{k-2}$ = q.litemset$_{k-2}$;// 下一步删除 $c \in C_k$ 所有序列, 且 c 的某些序列(k - 1) 就不在 L_{k-1} 中

for 所有 $c \in C_k$ 的序列 do

for 所有 c 的(k - 1) 序列 do

if ($s \in L_{k-1}$) then delete 来自于 C_k 的c

3.3 序列模式挖掘结果示例

如表 1 所示, 对其数据库进行序列挖掘。

表 1 用户行为时序数据库

CID	TID	ISet
1	2	*bcd*
2	1	*b*
2	2	*abc*
2	3	*bcd*
3	1	*ab*
3	2	*abc*
3	3	*bcd*

设 min_ support= 3, 序列模式的 L 长度为 3。表 2 为按照上述算法生成的频繁 1- 序列、2- 序列、3- 序列集合。

表 2 各频繁序列集

C_1		L_1		C_2		L_2		C_3		L_3	
1- 序列	支持度	1- 序列	支持度	2- 序列	支持度	2- 序列	支持度	3- 序列	支持度	3- 序列	支持度
a	3	*a*	3	*ab*	3	*ab*	3	*abc*	2	*bcd*	3
b	7	*b*	7	*ac*	2	*bc*	5	*abd*	0		
c	5	*c*	5	*ad*	0	*bd*	3	*acd*	0		
d	3	*d*	3	*bc*	5	*cd*	3	*bcd*	3		
				bd	3						
				cd	3						

序列挖掘的最后结果为一个频繁 3- 序列 *bcd*。

4 结束语

文中的创新之处就是提出了一个新的基于数据挖掘的入侵检测系统框架。在系统中, 将数据挖掘中的序列模式挖掘应用到入侵检测系统中来, 对其中数据挖掘的部分采用关联规则算法和序列模式挖掘算法相结合的方法。

参考文献:

[1] LEE Wenke. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems[D]. Columbia: Columbia University, 1999.

[2] 宋世杰. 基于序列模式挖掘的误用入侵检测系统框架研究[J]. 计算机工程与科学, 2006(1): 28- 30.

[3] Bace R G. Intrusion Detection[M]. US: Macmillan Technical Publishing, 1999.

[4] 李川川, 刘衍珩, 田大新. 基于序列模式挖掘的网络入侵检测系统[J]. 吉林大学学报, 2007(1): 121- 125.

[5] 钱 昱, 郑 诚. 基于序列模式的异常检测[J]. 微机发展, 2004, 14(9): 53- 55.

(上接第 141 页)

[2] Cortes C, Vapnik V. Support- vector networks[J]. Machine Learning, 1995, 20(3): 273- 297.

[3] Pdesa M J. 模式识别——原理、方法及应用[M]. 吴逸飞译. 北京: 清华大学出版社, 2002.

[4] 刘志刚, 李德仁, 秦前清, 等. 支持向量机在多类分类问题中的推广[J]. 计算机工程与应用, 2004(7): 10- 13.

[5] Arenas- Gsrcia J, Perez- Cruz F. Multi- class support vector machines: A new approach[C]// ICASSP, 2003. Hong Kong: [s.n.], 2003.

[6] Xu P, Chan A K. Support vector machines for multi- class signal classification with unbalanced samples[C]// Proceedings of the International Joint Conference on Neural Networks 2003. Portland: IEEE, 2003: 116- 119.

[7] Bottou L, Cortes C, Denker J, et al. Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition [C]// Proc of the Int Conf on Pattern Recognition. Jerusalem: [s.n.], 1994: 77- 87.

[8] Krebel U. Pairwise Classification and Support Vector Machines [C]// In: Scholkopf B, Burges C J C, Smola A J, eds. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: The MIT Press, 1999: 255- 268.

[9] 张爱丽, 刘广利, 刘长宇. 基于 SVM 的多类文本分类研究[J]. 情报学报, 2004(9): 6- 10.