基于超球支持向量机的兼类文本分类算法研究

秦玉平[1,2],王秀坤[1],李祥纳[2],王春立[1]
QIN Yu-ping[1,2],WANG Xiu-kun[1],LI Xiang-na[2],WANG Chun-li[1]

1.大连理工大学 电子与信息工程学院,辽宁 大连 116024
2.渤海大学 信息科学与工程学院,辽宁 锦州 121000
1.School of Electronic and Information Engineering,Dalian University of Technology,Dalian,Liaoning 116024,China
2.College of Information Science and Technology,Bohai University,Jinzhou,Liaoning 121000,China

QIN Yu-ping,WANG Xiu-kun,LI Xiang-na,et al.Study on multi-class text classification algorithm based on hyper-sphere support vector machines.Computer Engineering and Applications,2008,44(19):166-168.

Abstract: To multi-class text,a classification algorithm based on hyper-sphere support vector machines is proposed in this paper.Hyper-sphere support vector machine is used to get the smallest hyper-sphere in feature space that contains most texts of a class,which can divide the class texts from others.For the text to be classified,the distances from it to the centre of every hyper-sphere are used to confirm the classes that the text belongs to.The experimental results show that the algorithm not only has a faster performance on classification speed,but also has a higher performance on classification precision.
Key words: support vector machines;hyper-sphere;multi-class;classification

摘 要:针对兼类文本,提出了一种分类算法。对属于同一类别的文本,利用超球支持向量机在特征空间中求得一个能包围该类尽可能多文本的最小超球,使各类文本之间通过超球分隔开,达到分类效果。对待分类文本,计算它到各超球球心的距离,根据距离判定该文本所属的类别。实验结果证明,该算法不仅具有较快的分类速度,而且具有较高的分类精度。
关键词:支持向量机;超球;兼类;分类
DOI:10.3778/j.issn.1002-8331.2008.19.050 **文章编号**:1002-8331(2008)19-0166-03 **文献标识码**:A **中图分类号**:TP181

1 引言

随着计算机技术和通讯技术的飞速发展,人们可以获得的文本信息越来越多,对文本信息的有效组织和管理成为急待解决的问题。文本分类能够改善文本信息杂乱的状况,降低查询时间,提高搜索质量,快速有效地获取文本信息。因此文本自动分类技术越来越得到人们的关注。文本自动分类的任务是对未知类别的文本进行自动处理,判别它们所属预定义类别集中的一个或多个类别。基于机器学习的文本自动分类已经取得了很好的效果,提出了多种分类算法,如 k 最近邻算法、朴素贝叶斯算法、决策树算法、支持向量机等。支持向量机(Support Vector Machine,SVM)是建立在统计学习理论(Statistical Learning Theory,SLT)基础上的一种新的分类技术。它是基于结构风险最小化原则,根据有限样本信息在模型的复杂度和学习能力之间寻求最佳折衷,由于其出色的泛化性能,成为目前解决文本分类问题的主要工具[1,2]。

支持向量机本质上是二值分类器,并有很多优秀的训练算法。在处理多分类问题时,人们往往将其分解成一系列的二分类问题加以解决。常见的处理方法包括 1-a-r[3]、1-a-1[4]以及 DAGSVM[5]。但是这些算法都要求训练样本具有单一类别,并且只能为待分类样本标记一个类别。但兼类是文本的一个自然属性,也就是说一些文本不是确定的属于一个类别,而是多个类别。对于这种情况,前面提到的 SVM 多分类算法不能解决。本文提出了一种基于超球支持向量机的兼类文本分类算法(Multi-class Hyper Sphere SVM,MHSSVM),对于每个训练样本,可以属于多个类别;对于待分类样本,能够实现兼类标注。

2 超球支持向量机

设给定一类训练样本集 $\{x_i\}_{i=1}^{l}$ 和核函数 $K(x_i,x_j)$,其中,$x_i\in R^n$,K 对应某特征空间 Z 中的内积,即 $K(x_i,x_j)=\langle g(x_i),g(x_j)\rangle$,变换 $g:X\mapsto Z$ 是将样本从输入空间映射到特征空间。寻找特征空间的一个超球(a,R),其中 a 为球心,R 为球半径。超球应尽量包围样本的大部分映射,同时半径 R 应尽可能的小。当不存在偏远的点时,则寻找一个能够包围所有样本映射的最小超球;当存在偏远的点时,可以允许一部分样本影射在超球的外面,则寻找一个能够包围大多数样本映射的最小超球。当不知

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60603023);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2001CCA00700)。
作者简介:秦玉平(1965-),男,博士研究生,教授,主要研究领域为机器学习;王秀坤(1945-),女,博导,教授,主要研究领域为数据库系统;李祥纳(1982-),女,硕士研究生,主要研究领域为机器学习;王春立(1972-),女,博士,教授,主要研究领域为模式识别。
收稿日期:2007-09-27 **修回日期**:2007-11-30

道是否含有偏远的点时，通过引入一个非负松弛变量 $\xi_i, i=1,2,\cdots,l$，允许一部分样本的映射位于超球的外面。采用与寻找最优分类面类似的方法，通过求解带约束条件式(2)和式(3)的优化问题式(1)得到最小包围球[6-8]：

$$\min\ F(R,a,\xi_i)=R^2+\frac{1}{vl}\sum_i\xi_i \tag{1}$$

$$\text{s.t. } \|g(x_i)-a\|^2\leqslant R^2+\xi_i \quad i=1,2,\cdots,l \tag{2}$$

$$\xi_i\geqslant 0 \quad i=1,\cdots,l \tag{3}$$

其中，$0<v\leqslant 1$，用来控制超球的半径与它所能包围的样本数目之间的折衷。v 越小，惩罚也越大，对允许超球外面存在样本的约束程度也就越大。

为了求解上述优化问题，定义如下的 Lagrange 函数：

$$L(R,a,\beta,\gamma,\xi_i)=$$

$$R^2+\frac{1}{vl}\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\beta_i\{R^2+\xi_i-\|g(x_i)-a\|^2\}-\sum_{i=1}^{l}\gamma_i\xi_i \tag{4}$$

其中 $\beta_i\geqslant 0$ 为该类样本集的 Lagrange 系数。求解(4)的最小值，可以令该泛函对 R、a 及 ξ_i 求偏导，并令导数等于 0，得到：

$$\frac{\partial L}{\partial R}=2R\left(1-\sum_{i=1}^{l}\beta_i\right)=0\Rightarrow\sum_{i=1}^{l}\beta_i=1 \tag{5}$$

$$\frac{\partial L}{\partial\xi_i}=\frac{1}{vl}-\beta_i-\gamma_i=0\Rightarrow 0\leqslant\beta_i\leqslant\frac{1}{vl} \quad i=1,\cdots,l \tag{6}$$

$$\frac{\partial L}{\partial a}=-\sum_{i=1}^{l}2\beta_i(g(x_i)-a)=0\Rightarrow a=\sum_{i=1}^{l}\beta_i g(x_i) \quad i=1,\cdots,l \tag{7}$$

将约束条件式(5)~式(7)代入式(4)中，并进行合并整理，得到：

$$\max\ L(\beta_i)=L(R,a,\beta,\gamma,\xi_i)=\sum_{i=1}^{l}\beta_i K(x_i,x_i)-\sum_{i,j=1}^{l}\beta_i\beta_j K(x_i,x_j) \tag{8}$$

$$\text{s.t. } \sum_i\beta_i=1 \tag{9}$$

$$0\leqslant\beta_i\leqslant\frac{1}{vl} \quad i=1,2,\cdots,l \tag{10}$$

由式(7)可知，最小超球中心为带权系数 β_i 的线性加权组合：

$$a=\sum_i\beta_i g(x_i) \tag{11}$$

当 $\beta_i>0$ 时，对应的样本称为支持向量。当 $0\leqslant\beta_i\leqslant\frac{1}{vl}$ 时，对应的样本位于超球附近，任选其中一个该类样本 x 与球超球心之间的距离可确定超球半径。

$$R=\|g(x)-a\|=K(x,x)-2\sum_i\beta_i K(x,x_i)+\sum_{i,j}\beta_i\beta_j K(x_i,x_j)^{\frac{1}{2}} \tag{12}$$

当 $\beta_i=\frac{1}{vl}$ 时，对应的样本位于超球外面，称为野值或含噪声的样本。

3 兼类文本分类算法

设给定兼类样本集 $A=\{x_i,E_i\}_{i=1}^{l}$ 和核函数 $K(x_i,x_j)$，其中，$x_i\in R^n$，$E_i=\{y_{ij}\}_{j=1}^{p}$，$y_{ij}\in\{1,2,3,\cdots,N\}$，$N$ 是样本集 A 中含有的总类别数，$p(p\leqslant N)$是样本 x_i 的兼类数。K 对应某特征空间 Z 中的内积，即 $K(x_i,x_j)=\langle g(x_i),g(x_j)\rangle$，变换 $g:X\mapsto Z$ 是将样本从输入空间映射到特征空间。

设 A^m 为 A 中兼有类别 m 的样本子集，其中，$m=1,2,\cdots,N$。对于每一类样本集 A^m，利用超球支持向量机在特征空间确定一个超球(a_m,R_m)，其中，a_m 是该类超球的球心，R_m 为超球的半径。

对于待分类样本 x，根据公式(13)计算它到类 A^m 的最小包围球中心 a^m 的距离 $d_m(x)$，$m=1,2,\cdots,N$。根据 $d_m(x)$来判断 x 是否属于该超球所包含的类。

$$[d_m(x)]^2=\|g(x)-a_m\|^2=\|g(x)-\sum_i\beta_i^m g(x_i^m)\|^2=$$

$$K(x,x)+\sum_{i,j}\beta_i^m\beta_j^m K(x_i^m,x_j^m)-2\sum_i\beta_i^m K(x,x_i^m) \tag{13}$$

若 $d_m(x)>R_m$，$m=1,2,\cdots,N$，则根据式(14)计算样本 x 属于第 m 类的隶属度。

$$r_m=\frac{R_m}{d_m(x)} \tag{14}$$

根据式(15)确定 x 所属类别。

$$r=\max_m r_m \tag{15}$$

待分类样本 x 的分类过程具体描述如下：

步骤 1 根据公式(13)计算 $d_m(x)$，$m=1,2,\cdots,N$；

步骤 2 若存在 $d_m(x)\leqslant R_m$，则 x 所属于类别为$\{m|d_m(x)\leqslant R_m, m=1,2,\cdots,N\}$，转步骤 4，否则转步骤 3；

步骤 3 先根据公式(14)计算 r_m，然后根据式(15)计算 r，x 所属类别为$\{m|r_m\geqslant r, m=1,2,\cdots,N\}$，转步骤 4；

步骤 4 分类结束。

4 实验结果及分析

本文使用标准数据集 Reuters 21578，从中选取 5 类且一个文本所属类别最多为 3 的 808 篇文本进行实验分析。用其中的 539 篇文本作为训练样本，其余的 269 篇文本作为测试样本(见表 1)。将文本数据经过预处理后形成高维词空间向量，采用信息增益的方法来进行特征降维，向量中每个词的权重根据 tf-idf 公式计算。

实验中采用通用的准确率(AP)、召回率(AR)和平均 $F1$ 值

表 1 训练语料和测试语料

类别	corn	cotton	rice	soybean	wheat
训练集规模	168	44	44	79	204
测试集规模	84	22	22	40	101

(AF)作为评价指标。

$$\text{准确率}(P)=\frac{N_c}{N_a} \tag{16}$$

$$\text{召回率}(R)=\frac{N_c}{N_r} \tag{17}$$

$$F1=\frac{2*P*R}{P+R} \tag{18}$$

其中，N_c 代表对每个测试样本测试后得到的正确兼类数；N_a 代表对每个测试样本测试后得到的所有兼类数；N_r 代表每个测试样本实际的兼类数。

定义 1 平均准确率$(AP)=\dfrac{\sum P}{n}$ (19)

若 n 为测试样本总数，称为宏平均准确率(MAAP)；若 n 为兼类数相同的样本数，称为微平均准确率(MIAP)。

定义 2 平均召回率$(AR)=\frac{\sum P}{n}$ (20)

若 n 为测试样本总数,称为宏平均召回率(MAAR);若 n 为兼类数相同的样本数,称为微平均召回率(MIAR)。

定义 3 平均 $F1$ 值$(AF)=\frac{\sum F1}{n}$ (21)

若 n 为测试样本总数,称为宏平均 $F1$ 值(MAAF);若 n 为兼类数相同的样本数,称为微平均 $F1$ 值(MIAF)。

实验环境为 CPU Pentium 1.6 G,内存 512 M,操作系统 Windows XP。使用的核函数为径向基函数(Radial Basis Function,RBF)$K(x,y)=e^{-\gamma\|x-y\|^2}$,其中 $\gamma=0.01$,系统参数 $v=0.6$。算法实现参考了 Chang 和 Lin 所开发的 libsvm[9],并在此基础上进行了相应的修改。

表 2 和表 3 给出了对测试样本的分类结果。由表 2 和表 3 可以看出,MHSSVM 在保证单类文本分类精度的同时,实现了对兼类文本的分类,并且具有较好的准确率、召回率和 $F1$ 值。该算法不仅适应庞大的兼类识别问题,并且具有较强的扩展能力。

表 2 MHSSVM 的微平均准确率、微平均召回率和微平均 $F1$ 值

兼类数(样本数目)	1(212)	2(26)	3(2)
MIAP	71.94%	76.28%	66.67%
MIAR	75.47%	50.00%	66.67%
MIAF	73.11%	58.59%	66.67%

表 3 MHSSVM 的宏平均准确率、宏平均召回率和宏平均 $F1$ 值

MAAP	MAAR	MAAF
72.36%	72.64%	71.49%

5 结论

本文针对兼类文本的分类,提出了一种基于超球支持向量机的兼类文本分类算法。对具有同一兼类的文本,利用超球支持向量机在特征空间求得一个能够包围该类尽可能多映射的最小超球,使各类文本之间通过超球隔开,达到分类效果。每个超球的训练,只针对一类文本,因此,计算复杂度低,训练速度快。对待分类文本,计算它到各超球球心的距离,根据距离判定该样本所属的类别,分类简单快捷。实验结果证明,该算法不仅具有较快的训练速度,而且具有较高的分类速度和分类精度,是一种较为实用的兼类文本自动分类方法。

参考文献:

[1] Vapnik V.The nature of statistical learning theory[M].New York:Springer,1995.

[2] Joachims T.Text categorization with support vector machines:learning with many relevant feature[C]//Proceedings of 10th European Conference on Machine Learning,ECML-98.Berlin:Springer,1998:137-142.

[3] Bennett K P.Combining support vector and mathematical programming methods for classification[M].Advances in Kernel Methods:Support Vector Learning.Cambridge,MA:MIT Press,1999:307-326.

[4] Krebel U G.Pairwise classification and support vector machines[M].Advances in Kernel Methods:Support Vector Learning.Cambridge,MA:MIT Press,1999:255-268.

[5] Platt J C,Cristianini N,Shawe-Taylor J.Large margin DAGs for multiclass classification[M].Advances in Neural Information Processing Systems.Cambridge,MA:MIT Press,2000:547-553.

[6] 朱美琳,杨佩.基于支持向量机的多分类增量学习算法[J].计算机工程,2006,32(17):77-79.

[7] 张翔,肖小玲,徐光祐.基于样本之间紧密度的模糊支持向量机方法[J].软件学报,2006,17(5):951-958.

[8] 唐发明,王仲东,陈绵云.支持向量机多类分类算法研究[J].控制与决策,2005,20(7):746-749.

[9] Chang C C,Lin C J.LIBSVM:a library for support vector machines[J/OL].Journal of Machine Learning Research,2005,6:1889-1918.[2007-04].http://www.csie.ntu.tw/~cjlin/libsvm.

(上接 165 页)

gineering.Los Alamitos,CA:IEEE Computer Society Press,1999:49-71.

[2] Kaustubh R,Matti A,William H,et al.Automatic recovery using bounded partially observable markov decision processes [C]//Proceedings of the 36th International Conference on Dependable Systems and Networks,2006.

[3] Cassandra A,Nodine M,Bondale S,et al.Using POMDP-based state estimation to enhance agent system survivability[C]//Proceedings of 2005 IEEE 2nd Symposium,2005:11-20.

[4] Anthony R,Lcslic P,Michael L.Acting optimally in partially observable stochastic domains[C]//Proceedings of the 12th National Conference on Artificial Intelligence,Seattle,WA,1994.

[5] Cassandra A,Littman M,Zhan G,et al.Incremental pruning:a simple,fast,exact method for partially observable markov decision process[C]//Proceedings of the 3rd Conference on Uncertainty in Artificial Intelligence,San Mateo,1997:54-61.

[6] Aaron H,Karl K,Marshall B.A framework to control emergent survivability of multi agent systems[C]//Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems,2004:28-35.

[7] Holland H.Adaptation in natural and artificial systems[M].Ann Arbor,Michigan:University of Michigan Press,1975.

[8] Goldberg D,David E.Genetic algorithms in search,optimization and machine learning[M].Massachusetts:Addison-Wesley Publishing Company,1989.

[9] Goldberg D E,Lingle R.Alleles,loci,and the traveling salesman problem[C]//Proceedings of the 1st International Conference on Genetic Algorithms and Their Applications,1985:154-159.

[10] Davis L.Job shop scheduling with genetic algorithms[C]//Proceedings of the 1st International Conference on Genetic Algorithms and Their Applications.USA:Lawrence Erlbaum Associates,1985:136-140.

[11] Oliver L M,smith D J,Holland J R C.A study of permutation crossover operators on the traveling salesman problem[C]//Proceedings of the 2nd International Conference on Genetic Algorithms,Lawrence Erlbaum Associates,1987:224-230.

[12] Chen Y,Huang H,Jana R,et al.iMobile EE-an enterprise mobile service platform[J].ACM Journal on Wireless Networks,2003,9(4):283-297.