**融合数据分布特征的保序学习机**

张志剑，刘忠宝\*，李花，王文杰

###### (中北大学 软件学院 太原 030051)

摘要：如今支持向量机（Support Vector Machine，SVM）算法已经广泛应用于诸多实际问题中。但是在分类决策中仅依靠支持向量（Support Vector，SV）来确定最优超平面，对于众多非支持向量数据特征并没有合理的利用。

基于以上问题提出融合数据分布特征的保序学习机。引入类内离散度（Within-Class Scatter）表征数据内部的分布特征，让类内离散度最小来提高分类精度；通过各类样本数据中心相对位置不变保持了全局样本顺序不变；并证明本方法和核心向量机对偶形式等价解决了大规模分类问题。在人工数据集上的实验表明中小规模与大规模标注数据集均有较好表现。

关键词：类内离散度；支持向量机；全局保序；核心向量机

**Rank Preservation Learning Machine based on Data Distribution Fusion**

ZHANG Zhi-jian, LIU Zhong-bao

(SoftwareSchool, North University of China, Taiyuan 030051, China)

**Abstract：**

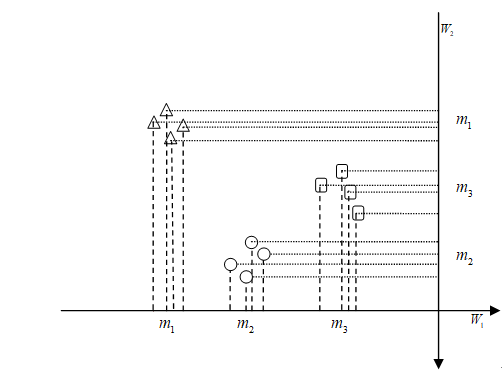
**Keywords：**Within-Class Scatter(*Sw*)；Support Vector Machine (SVM); global rank preservation；Core Vector Machine(CVM)

## 1引言

支持向量机（Support Vector Machine，SVM）由Vapnik和Corinna最早提出[1]，已经广泛应用于机器学习、数据挖掘、模式识别等领域。在解决小样本、非线性和高维度的模式识别中表现为速度快、精度高、理论支持好等优点[2]。SVM是基于结构风险最小化理论和VC（Vapnik-Cher-vonenkis）维理论，目标是通过在特征空间中寻找一个最优的超平面，从而得到全局最优解。在SVM提出后，众多学者提出了许多改进算法，包括（1）CSVM，常用的SVM称为标准SVM，也称为CSVM；（2）VSVM，在标准SVM中使边际最大而训练误差要求最小，其中参数*C*来调节两者的平衡，然而选取最优*C*是较为困难的，于是提出了VSVM[3]；（3）加权支持向量机（Weighted Support Vector Machine，WSVM），在实际问题中，不同样本在训练时权重是不同的。针对不同的样本应选取不同的惩因子，从而提出了WSVM；（4）最小二乘支持向量机（Least Squares Support Vector Machine，LSSVM），为解决标准SVM中不等式约束条件所带来的计算复杂，边界定义不清晰等问题，使用等式约束条件代替不等式约束条件，Suykens等人提出了LSSVM[4]；（5）拉格朗日支持向量机（Lagrange Support Vector Machine，LSVM），为了在处理大型线性数据集和中型规模非线性数据集时的收敛速度，提出了LSVM[5]；（6）核心向量机（Core Vector Machine, CVM），所以在训练大规模数据的分类器时会消耗大量时间和空间成本。Tsang等人提出了基于最小包含球(Minimum Enclosing Ball, MEB)的CVM[6]；（7）模糊支持向量机（Fuzzy Support Vector Machine，FSVM），在解决实际问题时，数据集中会存在不同程度的噪声，为解决噪声问题，Lin 等人提出了FSVM，通过引入模糊隶属度降低噪声和异常值对训练结果的影响[7]。

在标准支持向量机求解过程中，需要假设训练样本线性可分，才能继续求解。然而实际问题中的数据往往在原始空间是线性不可分的，所以引入了核函数（Kernel Function），通过核函数可以将数据从原始空间映射到高维空间中，使其线性可分。

上述几种方法在某些实际应用中均取有良好的使用效果，但仍然面临一些挑战：（1）分类过程并未考虑数据样本内部的分布特征，造成了数据资源的浪费，无法进一步提升分类性能；（2）分类结果忽视了各类样本的相对关系；（3）无法解决大规模分类问题。假设特征空间有三类样本，其先后顺序为*m1*、*m2*、*m*3，分类结果应保证三类样本的相对顺序不变。因此，三类样本投影在*W*1方向上优于*W*2方向(如图1所示)。鉴于此，提出融合数据分布特征的保序学习机（Rank Preservation Learning Machine based on Data Distribution Fusion, RPLM-DDF），该方法引入线性判别分析（Linear Discriminant Analysis, LDA）中的类内离散度*Sw*用以表征数据的分布特征。将各类样本中心相对关系考虑到最优化问题中，来确保分类过程中依然保持相对顺序不变。引入核心向量机来保证对于大型数据集依然有良好的可用性。



**图1 RPLM-DDF工作示意图**

## 2. RPLM-DDF

**2.1最优化问题**

假设样本集为，，，*X*表示所有样本特征的集合，代表第*i*类样本的均值。*Y*表示所有类别，代表第*i*类。类别数为*ｃ*，各类样本数为*Ni* (*i*=1,2,…,*c*)，*N*为样本总数。

融合数据分布特征的保序学习机（RPLM-DDF）使数据样本的类内离散度尽可能小，来提高分类精度；并通过各类样本的中心位置顺序不变来保持各类样本的顺序不变。则RPLM-DDF的最优化问题可描述为如下形式：

 （1）

 （2）

其中*W*为分类超平面的法向量，参数*β*为平衡因子，*ρ*为各类样本间距，*ν*是常数用来制约*ρ*，使得*νρ*达到最好的约束效果。 是各类样本的均值，*c*是类别数，**是类内离散度，其定义为：，其中 表示第*i*类样本集合，。

由Lagrangian定理可得：

 （3）

 （4）

 （5）

将（4）、（5）式带入（1）式中，去掉常数项可得如下对偶形式：

 （6）

 （7）

**2.2判别函数**

RPLM-DDF的判别函数为：

 （8）

其中

**2.3 时间复杂度分析**

RPLM-DDF方法的求解主要包含大小为矩阵的转置运算，其时间复杂度为；大小为 Hessian矩阵QP问题的求解运算,时间复杂度为。所以RPLM-DDF的时间复杂度为，但是，则RPLM-DDF的时间复杂度可用近似表示。

**2.4 非线性形式**

**2.4.1核化形式**

诸多实际问题中的数据在原始空间往往不是线性可分的，使用非线性映射函数将样本数据映射到高维空间中，使原本在低维空间线性不可分的问题，转化为高纬空间线性可分的问题。

假设映射函数满足条件时，RPLM-DDF最优化问题的非线性形式可表示为：

 （9） （10）

其中 ，

上述优化问题的核化对偶形式为：

 （11） （12）

**2.4.2核函数形式**

通过引入核函数，无需知道非线性变换函数的具体形式及参数，高维空间中的内积可以通过核函数直接运算，升维后算法复杂度也没有随着维度增加而增加[8]。但是使用核函数时由于是未知的，所以无法直接求解和，因此不能直接求解（11）式的对偶问题，故提出了一种方案解决上述问题，以下推论均假设只有两类数据：

原始问题转化为如下形式：

 （13） （14）

根据再生核Hilbert空间(RKHS)的性质，*W*可以写成[9]， 。

式（1）中:  （15）

其中*Y*是对角矩阵，，*G*是一个由核函数内积组成的矩阵， 

则可以表示为：

 （16）

其中是高维度特征空间中第*i*类的样本均值，表示为

取第一类为例：

 （17）

其中：；是阶单位矩阵；是阶填充的矩阵；*Y*是对角矩阵,定义为；



类似的可以得到第二类的表达式，所以:

 （18）

将（15）、（18）式带到（13）式可得：

 （19） （20）

令

可得：

 （21）

 （22）

由Lagrangian定理可得：

 （23）

*L*对*b*和*ρ*求偏导，令偏导等于0，得：

 （24）

 （25）

将（24）、（25）式带到（23）式中，可得：

 （26）

在对求偏导，可得：

 （27）

将(27)式带回(26)式中可得：



 （28）



 应该满足以下KKT条件：

 （29）

最终决策函数为：

 （30）

其中，。

**2.5大规模分类问题**

**2.5.1 核心向量机**

引入核心向量机保证RPLM-DDF对大规模分类问题依然适用。核心向量机QP问题的求解和计算几何中的最小包含球（Minimum Enclosing Ball，MEB）问题是等效的。通过有效的近似MEB算法，利用逼近率为的近似算法得到核心集（Core-set），而核心集的规模远远小于原始数据规模，从而降低了算法时间、空间复杂度，使SVM处理大规模数据成为可能。并且核心向量机可以与核函数共同使用，其时间复杂度是线性的，空间复杂度是独立的。对大规模数据集的实验表明，核心向量机与现有的SVM拥有相同的精度，但速度更快，可处理更大规模的数据集。核心向量机算法在处理大规模数据集时发挥了惊人的效率[10]。

**2.5.2 最小包含球**

最小包含球最优线性形式如下：  




其中为超球体的球心，为超球体的半径。

核化形式如下：





其中表示从低维空间到高维空间的映射。

由Lagrangian定理可得：





其中，核函数，，。由于等于常数并且，则上式最终可表达为：



**2.5.3 RPLM-DDF和最小包含球关系**

令 ，RPLM-DDF的QP问题可转换为：



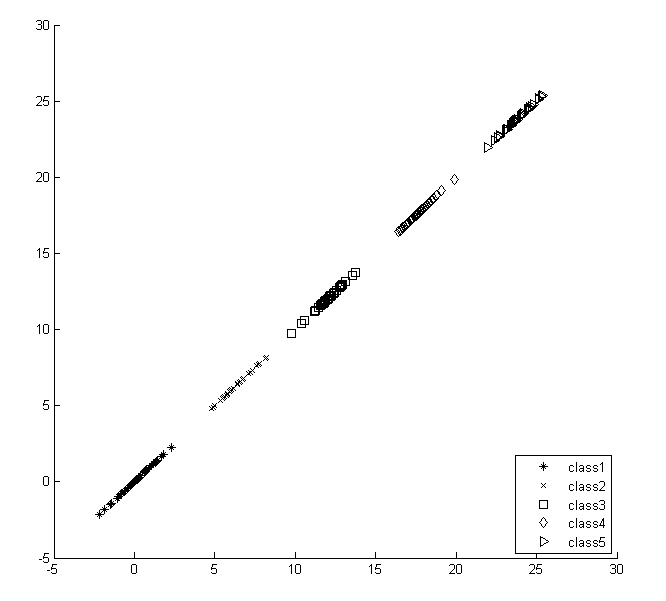
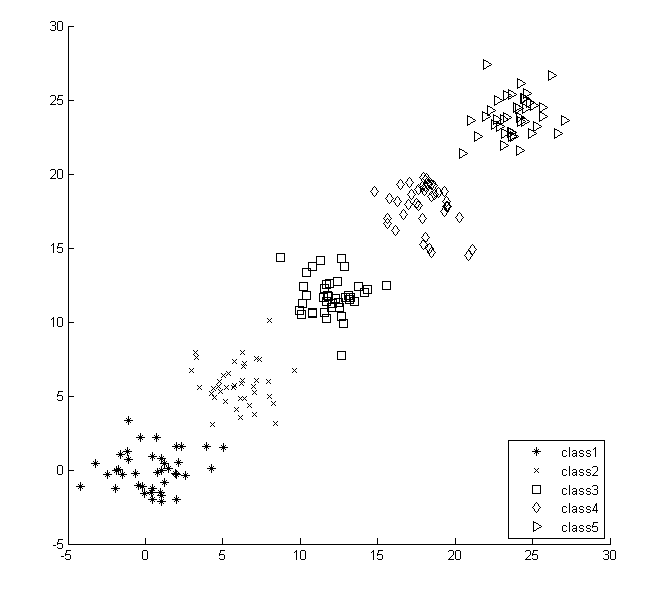


其中：，。RPLM-DDF形式和最小包含球形式等价，故RPLM-DDF可以使用MEB来解决大规模分类问题。

**3. 实验分析**

* 1. **人工数据集**

人工生成五类数据集，各类样本40个，各类中心点分别是(0,0)、(6,6)、(12,12)、(18,18)、(24,24)，标准差为2并服从Gaussian分布。生成数据集如图(a)所示，通过RPLM-DDF求得方向向量为***W***，将生成数据投影到***W***后得到(b)图。



**(a) 人工实验数据集 (b) 实验结果**

**图1 人工数据集及实验结果**

由图1可知，本文提出方法具有良好可分性，并且可以保持原始数据位置相对顺序不变。

**3.2中小型数据集**

本文实验采用的数据集如表1所示，实验过程中选取60%数据作为训练集，剩余数据作为测试集。

**表1 实验数据集**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Instances Number** | **Dimensions** | **Class Number** |
| **Iris** | **150** | **4** | **3** |
| **Liver** | **345** | **7** | **2** |
| **Glass** | **214** | **10** | **5** |
| **Wine** | **178** | **13** | **3** |
| **German** | **1000** | **20** | **2** |

实际问题中经常使用的核函数有：线性核函数（Linear Kernel Function），多项式核函数（Polynomial Kernel Function），高斯核函数（Gaussian Kernel Function），Sigmoid核函数。不同的核函数在不同应用环境下均有良好的性能[11]。但是在RPLM-DDF算法中因高斯核函数无论数据集规模大小，维度高低均适用，在低维线性不可分情况下有更为出色的表现，在选择最优参数情况下，分类效果最佳。对比了四种不同核函数的精度（参看图2），综上所述，本次实验中使用高斯核函数进行计算。

**图2 核函数与实验结果**

本文实验采用交叉验证的方法，有效地避免过拟合以及欠拟合的发生，最后得到的结果较有说服性[12]。将RPLM-DDF与SVC（Support Vectors Classification）、逻辑回归（Logistic Regression）、朴素贝叶斯（Naive Bayes）进行比较实验。使用网格搜索方法，在恰当的范围划分网格并遍历网格内所有点进行取值，得到参数[13]。*V*在{0.01,0.1,0.5,1,3,5,10}中选择，在{ }选择，‾*x*是训练样本平均范数的平方根。惩罚参数*C*在{0.01,0.05,0.1,0.5,1,5,10}中选择。实验参数如表2所示，实验结果如表3所示。

**表2 实验参数**

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | 参数 | | |
| SVC | 逻辑回归 | RPLM-DDF |
| Iris | =1=1 | C=1 | =,=0.1 |
| Liver | =0.01 C =0.5 | C=5 | =,=0.1 |
| Glass | =0.01 C =0.5 | C=10 | =,=0.5 |
| Wine | =0.01 C =1 | C=0.05 | =,=0.1 |
| German | =0.1 C =0.5 | C=1 | =,=0.1 |

**表3中小规模数据集对比实验结果**

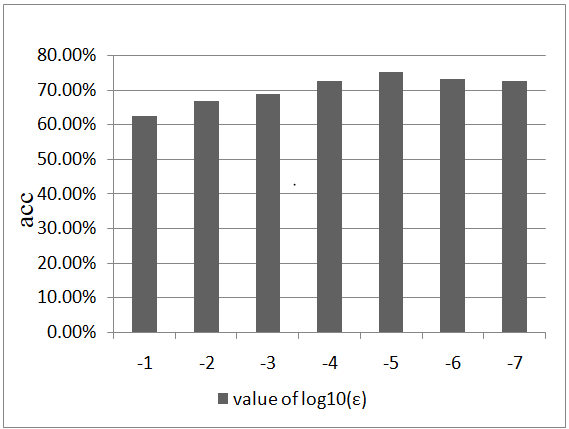
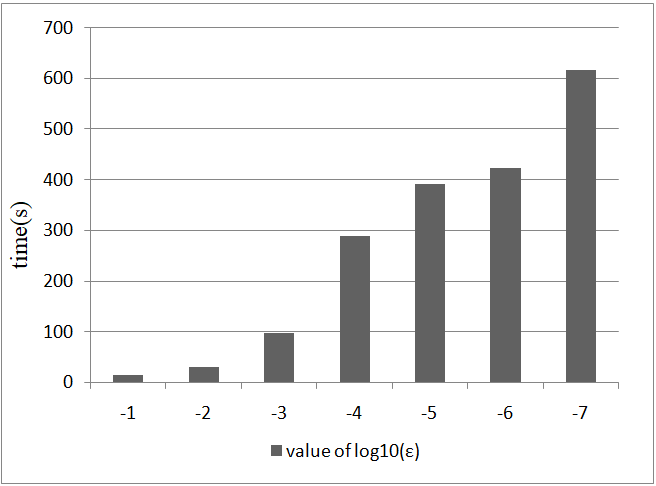
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Set | SVC | Logistic Regression | Naive Bayes | RPLM-DDF |
| Iris | 96.67% | 96.67% | 95.00% | **98.33%** |
| Liver | 64.49% | 65.21% | 60.86% | **70.28%** |
| Glass | 97.67% | 90.69% | 90.69% | **98.83%** |
| Wine | 65.27% | 97.22% | 93.26% | **98.61%** |
| German | 71.25% | 76.75% | 72.50% | **77.50%** |
| Average | 79.07% | 85.31% | 82.46% | **88.71%** |

由表3可知：在平均分类性能上RPLM-DDF相对于SVC、Logistic Regression、Naive Bayes具有更高的精度。在较为简单的分类任务（Iris数据集）中表现相当；在Liver、Glass、Wine、German数据集中较另外三种算法均有明显提升。故相比于传统算法RPLM-DDF能够更为出色的完成分类任务。

**3.3 大型数据集**

**3.3.1 Bank Marketing Data Set数据集**

实验采用Bank Marketing Data Set数据集，共有45211个样本，17维描述信息，分为两类。60%的数据集作为训练样本，剩余数据集作为测试样本。将在{}中选取。对实验时间影响如图3(a)所示，对实验精度影响如图3(b)所示。



(a) 和实验时间 (b) 和实验精度

**图3 对实验RPLM-DDF的影响**

**3.3.2 性能分析**

对数据集重新划分，将数据集的20%，40%，60%，80%作为训练集，并从剩余数据任取500个作为测试集。实验结果如表4所示,Acc表示正确率，单位为百分比(%)；Time表示训练时长，单位为秒(s)。

**表4 RPLM-DDF对大规模数据分类结果**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset size | Abalone | | Bank | | California | |
| Acc | Time | Acc | Time | Acc | Time |
| 20% | 61.46 | 80.12 | 63.68 | 156.32 | 46.03 | 243.84 |
| 40% | 70.21 | 130.45 | 67.32 | 278.53 | 54.58 | 403.47 |
| 60% | 75.36 | 173.26 | 71.58 | 295.72 | 60.26 | 672.94 |
| 80% | 76.14 | 197.63 | 77.04 | 331.18 | 64.57 | 734.28 |

由表4可知，随着训练样本的增加，RPLM-DDF分类精度呈上升趋势。但是这种上升并不是无限的，例如Abalone数据集由60%增长到80%时正确率并没有大幅增长。训练时间随着训练样本的增加而增加，但是RPLM-DDF能在有限的时间内高精度的完成分类任务。

**4．结论**

由于当前分类方法未考虑数据集中的类内结构，导致这种有价值的信息流失。鉴于此，提出融合数据分布特征的保序学习机（RPLM-DDF）。RPLM-DDF主要优势在于：（1）在考虑最优化问题时将类内结构融合起来，合理有效的利用这种信息，提高了算法分类精度；（2）较好的保持了数据的相对位置不变；（3）通过核心向量机的思想解决大规模分类问题。然而，RPLM-DDF方法仍然依赖参数的选取，如何更加高效的选择最优参数是下一步重点要进行的工作。

**参考文献**

1. 邓乃扬, 田英杰. 支持向量机:理论、算法与拓展[M]. 科学出版社, 2009.
2. WANG Hai-yan, LI Jian-hui, YANG Feng-lei. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5):1281-1286.
3. Crisp D J, Burges C J C. A Geometric Interpretation of v-SVM Classifiers[J]. Advances in Neural Information Processing Systems, 1999:244-250.
4. Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.
5. Mangasarian, O. L, Musicant, et al. Lagrangian support vector machines[J]. J.mach.learn.res, 2008, 1(3):161-177.
6. Tsang I W, Kwok J T, Cheung P M. Core Vector Machines: Fast SVM Training on Very Large Data Sets[M]. JMLR.org, 2005.
7. Lin C F, Wang S D. Fuzzy support vector machines[J]. IEEE Trans Neural Netw, 2002, 13(2):464-471.
8. 郭丽娟, 孙世宇, 段修生. 支持向量机及核函数研究[J]. 科学技术与工程, 2008, 8(2):487-490.
9. Wahba G. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV[C]. Advances in kernel methods. MIT Press, 1998:69-88.
10. 史荧中, 王士同, 王骏,等. 基于最小包含球的非静态大数据集的快速分类算法[J]. 控制与决策, 2013(7):1065-1072.
11. Müller K R, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2):181.
12. 王健峰, 张磊, 陈国兴,等. 基于改进的网格搜索法的SVM参数优化[J]. 应用科技, 2012, 39(3):28-31.
13. Wahba G, Lin Y, Zhang H. Generalized Approximate Cross Validation For Support Vector Machines, Or, Another Way To Look At. [J]. 1999.