第34卷 第20期 计 算 机 工 程 2008年10月
Vol.34 No.20 Computer Engineering October 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)20—0198—02 文献标识码：A 中图分类号：TP18

一种用于非平衡数据的SVM学习算法

蒋 莎，张晓龙
(武汉科技大学计算机学院，武汉 430081)

摘 要：在实际应用中的分类数据往往是非平衡数据，少数类别的数据可能有很大的分类代价。分类性能不仅要考虑分类精度，同时要考虑分类代价。该文扩展了支持向量机(SVM)学习方法，对于以高斯核为核函数时的少数类和多数类使用不同的惩罚参数C^+, C^-以获得高敏感度的超平面，并提出利用遗传算法对SVM的学习参数进行优化调整。给出一种新的评价函数，对分类结果的质量进行评价。实验结果证明，算法对于非平衡数据的分类有较好的效果，对少数类样本预测的准确性较高。

关键词：支持向量机；非平衡数据；评价函数；学习参数优化

SVM Learning Algorithm Used in Imbalance Data

JIANG Sha, ZHANG Xiao-long
(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081)

【Abstract】 In practice, training data is usually imbalanced, one class is "rare" relative to the other, and misclassification cost of the rare class may be much greater than the cost of the other class. In this situation, accuracy and the misclassification cost should be considered. This paper extends the Support Vector Machine(SVM) learning method, based on the Gauss kernel, by the use of C^+(the weight assigned to the rare class), and C^-(the weight assigned to the other class)to train more sensitive hyperplane, which is optimized by generic algorithm. Meanwhile, a new sensitive quality measure function is introduced in the optimization process. Experimental results show that the optimized algorithm has competitive performance when dealing with the rare class in the imbalance training data.

【Key words】 Support Vector Machine(SVM); imbalance data; measure function; learning parameters optimization

1 概述

使用高斯核函数的标准的支持向量机(Support Vector Machine, SVM)[1]对非平衡数据分类性能与典型分类器一样不能满足人们的需要[2]，为此用于对非平衡数据分类的SVM一般对数据集中不同的类别使用不同的惩罚参数C^+, C^-来代替原有的C参数，扩展成为多参数，使得分类时错分不同的类需要不同的代价来提高对少数类样本的分类能力。同时，对于非平衡数据的分类结果，一般使用交叉验证正确率或者F-measure进行评价，但是前者仅仅考虑了分类结果的整体正确率，对少数类的分类结果较为忽视；后者对少数类(正类)的分类精确度与敏感度进行折中，使得对于少数类的分类结果评价能同时兼顾精确度与敏感度，但是，F-measure对于敏感度与精确度的权重是相同的，使得在相当多的对少数类的分类结果的评价存在偏颇，使分类性能受到限制。

针对以上问题，本文对于非平衡数据的多数类和少数类在SVM中使用不同的惩罚参数C^+, C^-取代原有的C参数，以获得高敏感度的超平面，提出了一种新的评价函数，并使用遗传算法进行SVM算法参数选择来提高SVM对非平衡数据集的分类能力。

2 非平衡数据分类性能评价函数

2.1 评价函数 F-measure

分类结果的评价标准决定分类结果的优劣，并让SVM获得较优学习参数。分类的准确率是被正确分类的样本占数据集样本总数的比例，是分类任务中最经常使用的评估度量，它在对少数类进行度量时的缺点是显而易见的。相对于多数类来说，少数类样本对准确率的影响力较小，因为在不识别出任何少数类样本的情况下仍可以得到很高的正确率。

因此，在分类结果评价中必须考虑到正类(少数类)预测的正确率，在评价函数中考虑到错分正类的代价，而不仅仅是总体正确率。有一个使用较为广泛的评价函数，称为F-measure[3]。文献[3]中各参数定义如表1所示。

表1 F-measure中各符号

	实际的类别 $A+$	实际的类别 $A-$
预测的类别 $A+$	TP(正确的正类)	FP(错误的正类)
预测的类别 $A-$	FN(错误的负类)	TN(正确的负类)

因此，正类分类的正确率 P 可以定义为

$$P=\frac{TP}{TP+FP}\in[0,1] \tag{1}$$

正类分类的敏感度 R 定义为

$$R=\frac{TP}{TP+FN}\in[0,1] \tag{2}$$

由以上得F-measure评价函数为

$$F=\frac{2\times P\times R}{P+R}\in[0,1] \tag{3}$$

因而，分类结果最佳的情况是使评价函数值趋向最大值1。但F-measure存在的问题是仅有对正类分类结果的精确度和敏感度的一个固定的折中，对于需要构造高敏感度超平面的SVM来说是不够的，因此，需要构造一个具有更高敏感度的评价函数。

作者简介：蒋 莎(1982－)，女，硕士研究生，主研方向：机器学习；张晓龙，教授、博士

收稿日期：2007-12-13 **E-mail：**jiangsha107@126.com

2.2 一种新的敏感的评价函数F_u-measure

对式(3)进行变换，可以得到以下等式：

$$F=\frac{2\times P\times R}{P+R}=\frac{1}{\frac{1}{2}\times\frac{1}{R}+\frac{1}{2}\times\frac{1}{P}} \tag{4}$$

由此，F-measure 是对非平衡资料分类结果中正类分类的精确度和敏感度的一个等权重的折中。因此，可在式(4)中引入一个新的权值 μ， μ 的取值为[0, 1]，得到式(5)：

$$F=\frac{1}{\mu\times\frac{1}{R}+(1-\mu)\frac{1}{P}}=\frac{P\times R}{\mu P+(1-\mu)R}\qquad \mu\in[0,1] \tag{5}$$

当 $\mu=\frac{1}{2}$ 时，即为式(3)。由此，得到新的评价函数，称为F_u-measure，即式(6)。

$$F_\mu=\frac{P\times R}{\mu P+(1-\mu)R}\in[0,1],\ \mu\in[0,1] \tag{6}$$

很显然可以得到 F_μ 的值随 μ 的变化在 P 与 R 之间连续的平滑的变化，当 μ 越靠近 1 时，R 就具有越大的权重，反之，则 P 具有越大的权重；当 $\mu=0.5$ 时，式(6)就等同于式(3)。

使用式(6)评价分类结果时，可以根据需要对正类的错分赋予更大的代价，能够在 SVM 的模型选择中更好地指导 SVM，构造具有高敏感度的超平面，使 SVM 具有更好地对非平衡数据分类的性能。本文使用式(6)对于使用特定学习参数的 SVM 对非平衡数据分类结果进行评价，从而达到指导学习参数优化的目的，第 3 节中本文使用式(6)对学习参数优化过程中的交叉验证结果进行评价来得到优化的学习参数。

3 非平衡数据的学习参数优化

3.1 学习参数

本文选择目前使用最广泛的高斯核SVM。SVM的学习参数的选择与优化对于构造一个具有足够高敏感度的超平面来说是十分重要的，对于分类结果的影响很大[4]。已有的参数优化算法对多参数SVM的参数优化在时间开销上很大，在 3.2 节中使用遗传算法对参数进行优化，以得到一个较为快速的参数优化方法。

3.2 学习参数优化

遗传算法(Genetic Algorithm, GA)利用生物遗传学的观点，结合了适者生存和随机信息交换的思想，通过自然选择、交换、变异等作用机制，实现种群的进化。在寻优过程中，GA在解空间随机产生多个起始点并同时开始搜索，由适应度函数来指导搜索方向。是一种能够在复杂搜索空间快速寻求全局优化解的搜索技术。本文使用遗传算法对SVM的学习参数(C^+, C^-, σ)进行优化。

参数编码是实现GA的关键，采用何种方式编码是由问题本身的性质所决定的。对于SVM参数优化，是一个约束优化问题。对于约束优化问题，实值编码比二进制编码具有更高的搜索效率。因此，采用实值编码策略来实现SVM学习参数优化的编码。对于SVM学习参数优化的编码，本实验是每个个体由 3 个部分组成，分别为C^+, C^-, σ。使用评价函数式(6)作为目标函数，指定 μ 用来在一定程度上提高小比例样本分类敏感度所占权重，提高评价函数中小比例样本的分类结果对函数值的影响，增强SVM对非平衡数据分类能力。种群大小设置为 30，迭代次数为 100。具体算法如下：

输入：权值 μ，初始 C^+， C^-， σ。

输出：优化的学习参数 C^+, C^-， σ，K-折交叉验证结果 TP, FN。

遗传算法相关参数：种群规模 *Ps*，迭代次数 *Mg*，复制概率 *Cs*，交换概率 *Cc*，变异概率 *Cm*，搜索半径 *r*。

```
Procedure Optimazation( μ， C+， C−， σ )
Begin
  //设置遗传算法参数
  ChageGAConfig(Ps, Mg, Cs, Cc, Cm, r);
  //生成个体种群
  InitGenes();
  i = 0;
  //当前迭代次数 i < Mg 时
  While i < Mg then do
  Begin
      j = 0;
  //计算每个个体的适应性
      While j < Ps then do
      Begin
          //对每个个体进行 k-折交叉验证
          DoCrossValidation();
          //使用式(6)对分类结果进行评价
          Evaluate();
          //根据评价函数计算个体适应性
          Fitness();
      End
      Selection();   //选择操作
      Crossover();   //交叉操作
      Mutation();    //变异操作
  End
  //输出计算结果
  OutPutResualt( μ， C+， C−， σ , TP, FN)
End
```

对于参数C^+, C^-, σ，使用 30×30×30 格的网格法，需要训练 27 000 个SVM，而对于GA，只需要训练 30×100 个SVM，仅为网格法的 1/9。所以，相比传统的网格法，使用遗传算法对参数优化，尤其是多参数的SVM，有较大的速度优势。

4 实验结果

在扩充LIBSVM软件包的基础上，实现了本算法的实验平台。分类数据采用UCI中选出的breast-cancer, ionosphere数据集。这些分类数据在评价函数F_u-measure中，当 $\mu=0.5$ 时该函数等价于F-measure，因此在实验中，分别设置 $\mu=0.5$ 与 $\mu\neq0.5$ 两种值来对比F-measure与F_u-measure。使用以上 2 个不同评价函数的SVM同样都使用遗传算法来进行学习参数的优化。

breast-cancer, ionosphere数据集少数类与多数类的分布分别为 239/444, 126/225。实验中使用下式来确定F_u-measure中 μ 的取值：

$$\mu=\frac{N^-}{N^++N^-} \tag{7}$$

其中，N^+， N^- 表示样本中少数类、多数类样本的数量。

在实验中，也使用了双线性网格法(使用参数(C, σ))对以上 2 组数据进行实验，该方法使用交叉验证精确度来评价分类结果。因此，对于每组实验样本使用 3 种不同方法进行实验对比。最后使用F_u-measure对各组分类结果进行评价。

由实验结果表 2、表 3 可以看出，双线性网格法与使用F-measure、遗传算法优化学习参数的模型在整体错分率基本相当，然而前者在小比例样本分类错分数量除数据splice略高于后者之外，均低于后者，在小比例样本的分类能力上相对使用多参数(C^+,C^-, σ)并且使用F-measure指导训练的后者来说有较大的差距。 (下转第 202 页)