面向不平衡数据的模糊支持向量机

刘 凌,郭 剑,韩 崇
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:对于不平衡数据集,传统模糊支持向量机存在分类敏感等问题,且确定样本隶属度时大多只考虑距离因素,不能精准地反映样本点的重要程度,容易造成分类结果的偏差。文中提出一种改进的模糊支持向量机,在确定样本隶属度时,根据样本密度区分出不同类别的样本点,并分别赋予不同的隶属度值,提高了支持向量点的权重,降低了噪声点和孤立点对分类性能的影响。同时,进一步引入了不平衡类调节因子,以提高不平衡数据集的分类精度。实验结果表明,相比已有模糊支持向量机,该方法对于包含较多孤立点和噪声点的不平衡数据集具有更好的分类效果。

关键词:支持向量机;模糊支持向量机;不平衡数据集;样本密度
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2015)11-0038-06
doi:10.3969/j.issn.1673-629X.2015.11.008

Fuzzy Support Vector Machine for Imbalanced Data

LIU Ling, GUO Jian, HAN Chong
(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Traditional Fuzzy Support Vector Machines (FSVM) are sensitive to imbalanced data. They compute their fuzzy memberships mainly according to the factor of distance, which can not reflect the importance of the samples precisely and may lead to an error of classification results. To these problems, an improved FSVM is proposed in this paper. In the proposed FSVM, samples are firstly separated into different categories based on sample densities, and then they are assigned different fuzzy memberships. This method may improve the weight of support vectors and reduce the influence of outlier and noise points. Furthermore, the imbalanced factor is introduced to improve the classification precision of imbalanced data. The experimental results show that the improved FSVM has better performance for imbalanced data with more outlier and noise points.

Key words: support vector machine; FSVM; imbalanced data; sample density

0 引 言

支持向量机(Support Vector Machine, SVM)算法作为一种机器学习算法被广泛应用于统计分类以及回归分析之中[1-2],对于解决局部极小、维数灾难和过学习等问题具有较好的效果。但是,支持向量机也存在着训练时间长、受噪声点影响大、对不平衡数据集分类敏感等缺陷。有研究表明[3-4],当数据集不平衡时,SVM分类器对少数类的识别率很低;另外,当训练样本中含噪声点或者孤立点时,分类器可能会将噪声点误判断为支持向量,导致分类结果出现偏差。针对这些问题,Lin等[5-6]提出了模糊支持向量机,将模糊技术运用于支持向量机,为不同的样本赋予不同的权系数。Batuwita等[7-8]则在模糊支持向量机的基础上引入不平衡学习的方法,解决了模糊支持向量机对不平衡数据分类的敏感问题。Batuwita等将样本点之间的距离作为设计模糊隶属度函数的基础,但仅考虑了样本点线性可分的情况,在非线性可分时,并不能客观地反映样本点之间的位置关系。另外,仅基于距离的隶属度函数容易将对分类平面起决定性作用的支持向量点误判为噪声点或孤立点,造成严重的分类偏差。

针对上述问题,文中提出一种新的模糊支持向量机,通过样本紧密度区分出噪声点、孤立点,并赋予较

收稿日期:2015-01-04 **修回日期**:2015-04-15 **网络出版时间**:2015-11-04
基金项目:国家自然科学基金资助项目(61171053,61300239);教育部博士点基金资助项目(20113223110002);中国博士后科学基金资助项目(2014M551635);江苏省博士后科研资助计划项目(1302085B)
作者简介:刘 凌(1990-),女,硕士研究生,研究方向为大数据领域内的机器学习算法研究;郭 剑,副教授,硕士生导师,研究方向为无线传感器网络、无线多媒体传感器网络。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20151104.0952.042.html

小的隶属度值。对于边界样本与安全样本,结合考虑样本点与类中心的距离和样本点与分类超平面的距离,赋予边界样本更高的隶属度值,从而提高分类的准确率。同时,文中考虑了样本非线性可分的情况,准确描述了样本点之间的关系。通过实验仿真,该方法在噪声点、孤立点较多的情况下更能够准确地平衡少数类与多数类的分类精度。

1　模糊支持向量机

标准的支持向量机将所有样本点都看成是同等重要,对所有误分样本点都赋予相同的惩罚因子,但在现实应用场景中,训练样本点所属类别并不能明确给出,并且由于样本中噪声点和孤立点的存在,不同的样本点对分类超平面的作用是不同的,若不区分对待,将会导致严重的分类误差。因此,将含有重要意义的样本正确分类,并且忽略孤立点、噪声点对分类的影响是十分有必要的。模糊支持向量机在支持向量机的基础上,根据训练样本在训练过程中所起的作用不同,为每个训练样本赋予不同的隶属度,亦即为每个样本点赋予一个权值,再利用带有隶属度的支持向量机模型对样本集进行训练构造新的分类器,从而提高算法抵抗噪声点或孤立点的能力,提高分类精准度。

用模糊支持向量机进行分类,即选择一个适当的隶属度函数,根据隶属度函数得到每个训练样本点 (x_i,y_i) 的隶属度值 $u_i(0 < u_i \leqslant 1)$,u_i 即表示样本 x_i 对类别 y_i 的从属程度。于是训练集就变成了模糊化训练样本集 $S = \{(x_i,y_i,u_i),i = 1,2,\cdots,N\}$。其中,$x_i \in R^n$ 是训练样本集,$y_i \in \{-1,1\}$ 是样本标签,$0 < u_i \leqslant 1$ 是模糊隶属度函数取值范围。则求解最优超平面的优化问题则变为:

$$\min \frac{1}{2} \| \omega \|^2 + C\sum_{i=1}^{N} u_i\xi_i \tag{1}$$

$$\text{s.t.}\quad y_i(\omega \cdot \Phi(x_i) + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1,2,\cdots,N$$

其中,C 为惩罚因子且满足 $C > 0$;ξ_i 表示松弛变量。

由式(1)可以看出,每个样本点 x_i 所对应的模糊隶属度 u_i 被嵌入了目标函数当中。因此,$C \cdot u_i$ 表示对错分点的重视程度,$C \cdot u_i$ 越小,则损失参数 ξ_i 对目标函数值的影响越小,样本点 x_i 越不重要。相反,$C \cdot u_i$ 越大,对应样本越重要,x_i 被错分的概率就越小。由此可知,应尽可能减小噪声点或孤立点的 $C \cdot u_i$ 值,减小此类样本点对分类超平面的影响。

该问题可以转换为其对偶问题进行求解,其对偶问题为:

$$\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) \tag{2}$$

$$\text{s.t.}\quad y_i(\omega \cdot \Phi(x_i) + b) - 1 \geqslant 0$$

$$\sum_{i=1}^{n} y_i\alpha_i = 0$$

$$0 \leqslant \alpha_i \leqslant \mu_i C, i = 1,2,\cdots,N$$

由式(2)可以看出,模糊支持向量机与标准支持向量机的区别仅在于对偶问题中变量 α_i 的上界约束是随模糊隶属度函数 u_i 的变化而变化的,这也相当于为每个样本点分别赋予一个惩罚因子 $C \cdot u_i$。因此模糊隶属度 u_i 的确定成为决定模糊支持向量机工作性能好坏的关键。

2　模糊隶属度函数

在模糊支持向量机中,模糊隶属度函数的设计对于模糊支持向量机来说起到了至关重要的作用,它决定了某个样本点的重要程度。不同的隶属度函数会对数据处理结果以及算法复杂度产生不同的影响,而对于模糊支持向量机的分类问题,最优分类面主要由支持向量点决定。支持向量点位于类边缘,噪声点、孤立点也大多位于类边缘,对于不平衡数据集,经常会出现少数类中的样本点被误认为是噪声点、孤立点的现象。目前,存在很多方法用以构造模糊隶属度,但仍未形成一个统一的准则。在实际问题中,针对不同问题,设计的隶属度函数也不尽相同,现有的大部分隶属度函数都是基于距离因素的,这种方法可以通过样本点距离类中心的远近来判断一个样本点的重要程度。然而,这种方法有时无法将孤立点、噪声点从有效样本集中区分出来,以致赋予孤立点、噪声点错误的隶属度造成分类结果出现偏差。因此,如何设计出高性能的隶属度函数对模糊支持向量机的分类是至关重要的,设计的隶属度函数必须能够准确、客观地体现样本的重要程度。

样本点到类中心的距离是衡量样本重要程度的依据之一。文献[5]采用了基于距离的隶属度函数,将样本的隶属度值看作是样本与其所在类中心之间距离的线性函数。文献[9]使用 S 型隶属度函数,将样本到所在类中心的距离看作一个非线性关系。这两种方法都按照样本与类中心的距离准则考虑样本的隶属度,但是,往往无法区别支持向量点和噪声点。文献[8]给出了几种常用的隶属度函数:

$$f_{\text{lin}}^{\text{cen}}(x_i) = 1 - \frac{d_i^{\text{cen}}}{\max(d_i^{\text{cen}}) + \Delta} \tag{3}$$

$$f_{\text{exp}}^{\text{cen}}(x_i) = \frac{2}{1 + \exp(\beta d_i^{\text{cen}})}, \beta \in [0,1] \tag{4}$$

$$f_{\text{lin}}^{\text{hyp}}(x_i) = 1 - \frac{d_i^{\text{hyp}}}{\max(d_i^{\text{hyp}}) + \Delta} \tag{5}$$

$$f_{\text{cen}}^{\text{hyp}}(x_i) = \frac{2}{1 + \exp(\beta d_i^{\text{hyp}})}, \beta \in [0,1] \tag{6}$$

但这些隶属度函数也存在较多不足：

(1)文献[8]在计算距离时，只考虑样本线性可分的情况，而大部分样本都是非线性可分的，需要将高维度样本点映射到特征空间加以计算，使用核函数来表示样本之间的距离。

(2)若仅采用式(3)或式(4)来衡量一个样本的重要性，降低远离类中心但距离分类面很近的样本点的作用。如图 1 所示，分类超平面附近的样本点 A，B，C，D 距离类中心很远，因而隶属度值较小。但它们位于分类超平面附近，因此很有可能是支持向量点，对分类起着决定性作用。另外，样本点 A_1 与 A_2，B_1 与 B_2 都位于分类超平面附近，对于分类面的贡献是相近的，但它们与类中心的距离完全不同，导致隶属度值完全不同。

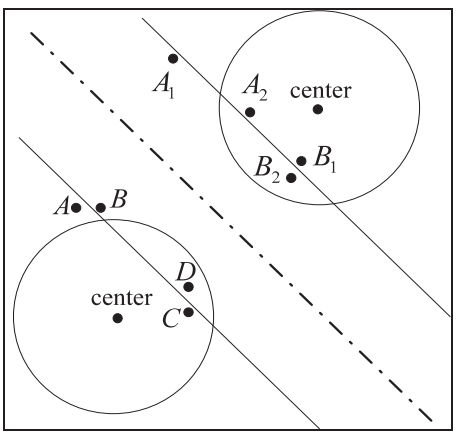

图 1 样本点距类中心示意图

(3)若仅使用式(5)、(6)，一些噪声点易被当作重要样本点赋予较高的隶属度值。如图 2 所示，样本点 A 距离分类超平面较近，因此被赋予较高的隶属度值，但样本点是噪声点，应该被忽视，即赋予极小的隶属度值，因此仅使用基于距离的隶属度函数，会使分类结果存在较大偏差。

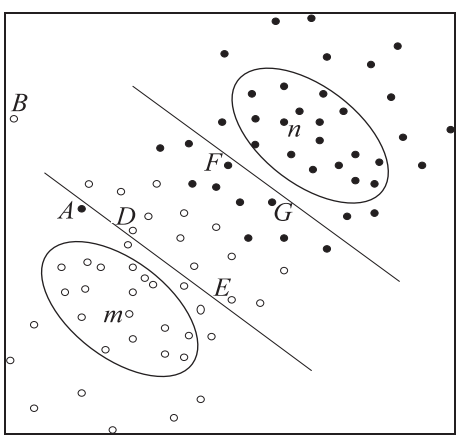

图 2 隶属度函数设计示意图

3 面向不平衡数据的模糊支持向量机

模糊支持向量机很好地提高了分类器的性能，但对于不平衡分类问题依然敏感。为了减小不平衡数据集带来的影响，文中使用 Different Error Costs (DEC) 算法[10]解决不平衡分类问题。同时设计了一种基于样本紧密度的隶属度函数，根据同类样本点与异类样本点的密度关系，有效地区分出噪声点与孤立点，降低了噪声点、孤立点对分类器的影响。

3.1 改进模糊支持向量机的基本框架

DEC 算法的思想与模糊隶属度函数类似，即通过给多数类样本和少数类样本分别赋予不同的惩罚因子。假设少数类为正类，其误分代价为 C^+，多数类为负类，其误分代价为 C^-。在进行不平衡调节时，使 $C^+ > C^-$，通过给少数类样本点赋予更大的惩罚因子 C^+ 来突出少数类样本的重要性，从而降低不平衡率对支持向量机分类器偏向性的影响。加入误分代价后的模糊支持向量机目标函数表达式如下：

$$\min\left(\frac{1}{2}\|\omega\|^2 + C^+ \sum_{[i=1 \mid y_i=+1]}^{N} u_i\xi_i + C^- \sum_{[i=1 \mid y_i=-1]}^{N} u_i\xi_i\right) \tag{7}$$

$$\text{s.t.} \quad y_i(\omega \cdot \Phi(x_i) + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1,2,\cdots,N$$

其中，C^+，C^- 分别为正、负类样本点的误分代价。

R. Akbani 等[11]已通过实验证明，当 C^-/C^+ 的比值等于 N^+/N^- 时(其中，N^+、N^- 分别为少数类、多数类样本数目)，DEC 算法效果最佳。$C^+ = C \cdot r^+$，$C^- = C \cdot r^-$，其中 C 为常数。根据 DEC 算法，设 $r^+ = 1$，$r^- = r$，r 即为正类样本数 N^+ 与负类样本数 N^- 的比例。u_i 为样本隶属度函数。因此，正类样本点隶属度范围为(0，1]，负类样本点隶属度范围为(0，r]。根据拉格朗日函数和 KKT 条件，求得其对偶问题为：

$$\max W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C^+\mu_i, y_i = 1, i = 1,2,\cdots,N^+$$

$$0 \leqslant \alpha_i \leqslant C^-\mu_i, y_i = -1, i = 1,2,\cdots,N^-$$

3.2 基于样本紧密度的隶属度函数

通过样本紧密度，文中将样本点分为四类，即噪声样本、孤立样本、边界样本以及安全样本。如图 2 中，A 为噪声样本，B 为孤立样本，簇 m、n 内的样本点为安全样本，D、E、F、G 为边界样本。文中方法利用样本密度的特性检测出噪声样本和孤立样本，为其赋予较小的隶属度值。而根据安全样本离类内中心较近，边界

样本离分类超平面较近的特性,对安全样本和边界样本采用距类内中心的距离与距分类超平面距离加权的方法为其赋予一定的隶属度值。

(1)噪声点、孤立点的检测。

在一个样本点的某个领域范围内,如果同类样本点数目越多异类样本点越少,则这个样本点很有可能是正常样本;当一个样本点领域范围内同类样本与异类样本数目都较少,则这个样本点很有可能是孤立点;而当一个样本点领域范围内异类样本较多而同类样本较少时,这个样本点则很大可能为噪声点。根据这一结论,则可以检测出噪声样本、孤立样本点,并为其赋予较小的隶属度值。为了描述文中方法,首先给出如下定义。

定义 1 平均距离:设有 n 个样本点,n 个对象两两之间的距离为 d_{ij},其中有 $d_{ij}=\|x_i-x_j\|$,当样本点非线性可分时 $d_{ij}=\|\Phi(x_i)-\Phi(x_j)\|$。则可定义距离矩阵:

$$\boldsymbol{R}=\begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n} \end{bmatrix} \tag{9}$$

所有对象的平均距离为:

$$D=\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{i,j}}{C_n^2} \tag{10}$$

定义 2 最大样本距离:

$$d_{\max}=\mathrm{Max}(d_{ij}),i,j=1,2,\cdots,n \tag{11}$$

定义 3 样本邻域:以样本点 x_i 为球心,r 为半径的一个球形范围。

根据文献[12]中得出的结论,当 r 取 $\frac{\sqrt{3}}{2}D$ 时,能够覆盖 x_i 领域内的大部分点。

定义 4 正密度 $\rho^+(x_i)$:样本点 x_i 在领域范围内同类样本点的个数,即:

$$\rho^+(x_i,r)=\left|\{x_j\,|\,D(x_j,x_i)\leqslant r,y_j=y_i\}\right| \tag{12}$$

定义 5 负密度 $\rho^-(x_i)$:样本点 x_i 在领域范围内异类样本点的个数,即:

$$\rho^-(x_i,r)=\left|\{x_j\,|\,D(x_j,x_i)\leqslant r,y_j\neq y_i\}\right| \tag{13}$$

定义 6 平均密度 $\bar{\rho}$:

$$\bar{\rho}=\frac{n}{t\left(\frac{\sqrt{3}}{2}\times d_{\max}\right)^m} \tag{14}$$

其中,n 为数据集中样本点的总个数;m 为样本点的维度;t 为常系数。

定义 7 样本密度 $\rho(x_i,r)$:

$$\rho(x_i,r)=\frac{n_i}{\left(\frac{\sqrt{3}}{2}aD\right)^m} \tag{15}$$

其中,n_i 是样本点领域范围内的样本数目;a 为可调参数,取值范围为(0,1]。当 a 取 1 时,$n_i=\rho^+(x_i)+\rho^-(x_i)$。

有了平均密度 $\bar{\rho}$ 和样本密度 $\rho(x_i,r)$ 之后,比较这两个密度即可以检测出孤立点。当 $\rho(x_i,r)<\bar{\rho}$ 时,样本点 x_i 为孤立点;当 $\rho(x_i,r)\geqslant\bar{\rho}$ 时,样本点 x_i 为非孤立点。此外,又可以通过正密度 $\rho^+(x_i)$、负密度 $\rho^-(x_i)$ 区分出非孤立点中的噪声点,当 $\frac{\rho^+(x_i)}{\rho^+(x_i)+\rho^-(x_i)}\leqslant\frac{\rho^-(x_i)}{\rho^+(x_i)+\rho^-(x_i)}$ 时,样本点为错分点,否则样本点为正常点。对于孤立点和错分点,分别赋予一个较小的隶属度值 δ,而对于正常点,其距离类内中心点越近,越有可能被正确分类,但距离类内中心远的点也有可能是对于分类超平面起重要作用的支持向量点。因此,对于正常点,采用基于样本点与类内中心的距离和基于样本点与分类超平面距离加权的方式为其分配模糊隶属度值。

(2)边界样本、安全样本隶属度的计算。

样本点离类内中心越近,越有可能被正确分类,应对其赋予越高的隶属度,其隶属度函数如式(3)、(4)。两式均是基于样本点到类中心的距离 d_i^{cen} 的,区别在于两个函数式对于 d_i^{cen} 的衰减方式不同。设 $\Phi(x_{\mathrm{cen}}^+)$ 为映射到特征空间的正类样本中心,$\Phi(x_{\mathrm{cen}}^-)$ 为映射到特征空间的负类样本中心。则有:

$$(d_i^{\mathrm{cen}})^+=\|\Phi(x_i)-\Phi(x_{\mathrm{cen}}^+)\|=\sqrt{\Phi^2(x_i)-\frac{2}{n^+}\sum_{x_j\in S^+}\Phi(x_j)\Phi(x_i)+\frac{1}{(n^+)^2}\sum_{x_j\in S^+}\sum_{x_k\in S^+}\Phi(x_j)\Phi(x_k)}=\sqrt{K(x_i,x_i)-\frac{2}{n^+}\sum_{x_j\in S^+}K(x_j,x_i)+\frac{1}{(n^+)^2}\sum_{x_j\in S^+}\sum_{x_k\in S^+}K(x_j,x_k)} \tag{16}$$

同理可得:

$$(d_i^{\mathrm{cen}})^-=\sqrt{K(x_i,x_i)-\frac{2}{n^-}\sum_{x_j\in S^-}K(x_j,x_i)+\frac{1}{(n^-)^2}\sum_{x_j\in S^-}\sum_{x_k\in S^-}K(x_j,x_k)} \tag{17}$$

但是根据文献[13]所述,离类中心距离远的点也有可能是对分类超平面起决定性作用的支持向量点,因此,单纯地根据样本点距离类中心的距离并不能很好地判断样本的重要性。文中同时考虑样本点与实际分类超平面的距离作为判断安全样本重要程度的标准。离分类超平面越近的点,越有可能是支持向量点,其重要程度越高,模糊隶属度越大;反之,离分类超平

面越远的点,模糊隶属度越小。则对于安全样本、边界样本,其模糊隶属度即为 $\zeta f_{\text{lin}}^{\text{cen}}(x_i) + \tau f_{\text{lin}}^{\text{hpy}}(x_i)$ 或 $\zeta f_{\text{lin}}^{\text{cen}}(x_i) + \tau f_{\text{exp}}^{\text{hpy}}(x_i)$。

(3)模糊隶属度函数设计。

根据(1)、(2)可得出样本点 x_i 的模糊隶属度函数 u_i 设计方法如下:

Input:样本点 x_i,δ

Compute $\rho(x_i, r)$

If $\rho(x_i, r) < \bar{\rho}$, $u_i = \delta$

Else

Compute $\rho^+(x_i)$, $\rho^-(x_i)$

If $\dfrac{\rho^+(x_i)}{\rho^+(x_i) + \rho^-(x_i)} \leqslant \dfrac{\rho^-(x_i)}{\rho^+(x_i) + \rho^-(x_i)}$, $u_i = \delta$

Else $u_i = \zeta f_{\text{lin}}^{\text{cen}}(x_i) + \tau f_{\text{lin}}^{\text{hpy}}(x_i)$ Or $u_i = \zeta f_{\text{exp}}^{\text{cen}}(x_i) + \tau f_{\text{exp}}^{\text{hpy}}(x_i)$

(其中 ζ、τ 为两个参数,当 $\zeta + \tau = 1$ 时,$u_i \in (0,1]$)

Output: u_i

由上面可以看出,文中方法具有如下优势:

(1)孤立点和错分点通过密度法被很好地检测出,并赋予较小的隶属度值,使其不影响分类结果。

(2)通过同时考虑样本点距离类中心与距离分类超平面的距离,使得支持向量样本点被赋予更大的隶属度值,突出了其分类过程中的重要程度。

(3)对少数类数据赋予较高的隶属度值,提高其重要程度。

4 实验结果与分析

为了验证文中提出方法的有效性和普适性,进行了如下实验。使用了 5 种失衡程度不同的 UCI 数据集作为验证集,这些数据集中很有可能存在一些噪声点或孤立点。表 1 按照样本正负比例递减的顺序给出了数据集的详细信息,其中包括正类样本数目(少数类)、负类样本数目(多数类)、样本总数、正负类不平衡比例、正类的序号。而对于原本是多类的数据集,选择其中一个类作为正类,其余剩下的所有类作为负类。

表 1 不同比例 UCI 数据集

数据集	正类数	负类数	样本总数	不平衡率	正类类号
Spambase	1 813	2 788	4 601	1:1.54	1
Haberman	81	225	306	1:2.78	2
Ecoli	77	259	336	1:3.36	2
Page-block	115	5 358	5 473	1:46.6	5
Unbalanced	12	844	856	1:70.3	1

面向不平衡数据集的评价标准很多,大部分是基于混淆矩阵的,如表 2 所示。

表 2 混淆矩阵

	正类(预测)	负类(预测)
正类(实际)	TP	FN
负类(实际)	FP	TN

其中,TP、FN、TN 以及 FP 依次代表分类正确的正类样本、假的负类样本、正确的负类样本以及假的正类样本的数目[14],且满足 TP+FN = N^+、TN+FP = N^- 以及 $N^+ + N^- = N$。常用的评价指标准确率(Accuracy)= $\dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$,但是对于高度不平衡数据集,该评价指标显然不能很好地反映一个分类器的性能。文中采用一种专门针对不平衡数据的评价指标[4],见式(18)、(19)、(20)。

$$S_{\text{e}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

$$S_{\text{p}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{19}$$

$$G_{\text{m}} = \sqrt{S_{\text{e}} \times S_{\text{p}}} \tag{20}$$

其中,S_{e} 代表正确分类正类的能力;S_{p} 代表正确分类负类的能力。由定义可知,S_{e} 和 S_{p} 越大越好,但在许多情况下,高 S_{e} 不一定具有高 S_{p}。因此引入 G_{m},G_{m} 是 S_{e} 和 S_{p} 的几何平均,G_{m} 越大,分类效果越好。

本实验使用 Weka 与 Libsvm 工具包作为模糊支持向量机的开发基础,实验前首先将数据进行归一化处理,归一后的范围为[0,1],实验中的核函数采用广泛应用的 RBF 核函数 $K(x_i, x_j) = \text{e}^{-\gamma(\|x_i - x_j\|^2)}$。为了更好地进行实验性能的对比,惩罚参数 C 和核参数 γ 采用五折交叉验证和网格搜索相结合的方法对其择优选择。其他参数设置为:$t = 1$,$a = 1$,$\Delta = 10^{-6}$,β 在 $\{0.1, 0.2, \cdots, 1.0\}$ 之间寻找最优范围。实验步骤具体如下:

Step1:先使用标准支持向量机对不平衡数据集进行训练,通过使用上文中的参数选择过程方法得到最优参数对 (C_1, γ_1),保存训练模型,得到标准支持向量机训练下的实验结果。

Step2:根据步骤 1 中的模型以及样本集,计算出每个样本的模糊隶属度值 u_i,根据正负类样本数目,计算出正负类样本对应的不平衡因子 C^+、C^-。

Step3:将模糊隶属度值与不平衡因子加入原来标准支持向量机模型中,在参数选取准则下得到最优参数对 (C_2, γ_2) 以及实验结果。

使用文献[7]所述的 FSVM - $\text{CIL}_{\text{lin}}^{\text{cen}}$、FSVM - $\text{CIL}_{\text{exp}}^{\text{cen}}$、FSVM - $\text{CIL}_{\text{lin}}^{\text{hpy}}$、FSVM - $\text{CIL}_{\text{exp}}^{\text{hyp}}$ 四种方法对样本

集进行训练与预测,并与文中算法相比较,实验结果如表 3 所示。

表 3　实验比较结果　　%

DataSet	Results	SVM	SVM CIL	FSVM CIL_{lin}^{cen}	FSVM CIL_{exp}^{cen}	FSVM CIL_{lin}^{hyp}	FSVM CIL_{exp}^{hyp}	Proposed CIL_{lin}	Proposed CIL_{exp}
Spambase	S_e	88.86	88.57	90.24	88.53	90.62	89.30	91.01	91.20
	S_p	96.09	95.66	95.91	93.26	95.59	93.76	95.92	95.34
	G_m	92.39	92.05	93.03	90.86	93.07	91.50	93.43	93.25
Haberman	S_e	33.31	54.56	42.53	43.78	48.25	25.48	45.09	47.87
	S_p	80.89	61.78	88.34	81.26	70.28	91.13	90.11	84.39
	G_m	51.91	58.06	61.30	59.65	58.23	48.19	63.74	63.56
Ecoli	S_e	80.75	96.17	88.91	92.34	87.51	90.93	91.55	92.73
	S_p	94.97	84.15	88.62	89.10	86.83	87.31	89.56	90.06
	G_m	87.57	89.96	88.76	90.71	87.17	89.10	90.55	91.39
Page-block	S_e	62.42	81.67	88.37	83.89	90.91	92.15	94.52	94.87
	S_p	99.68	83.05	87.22	88.13	89.06	90.27	93.84	94.91
	G_m	78.93	82.36	87.79	85.98	89.98	91.21	94.18	94.89
Unbalanced	S_e	23.33	70.00	58.37	53.89	47.83	62.15	64.87	74.52
	S_p	98.82	75.60	67.22	88.13	79.06	78.27	80.91	83.84
	G_m	48.02	72.74	62.64	73.23	61.49	69.75	72.45	79.04

从实验结果可以看出,对于不平衡数据集,正类(少数类)的分类正确率通常比负类(多数类)分类正确率小,且随着不平衡比例的增加,这种趋势表现的愈明显。与标准支持向量机相比,加入了不平衡因子的算法,其 S_e 值均有了明显提高,但因为一些噪声点和孤立点的影响以及各个数据集分布的不同,文献[8]中提出的模糊支持向量机分类效果并不一定都优于标准支持向量机。然而文中提出的模糊支持向量机在各个数据集上都具有较优的 G_m 值。对于 Spambase、Haberman、Ecoli 这些不平衡率较小的数据集,文中方法的实验结果与文献[7]中提出的 FSVM 实验结果相差不大,而对于 Pageblock、Unbalanced 这两个高不平衡率的数据集来说,文中方法则具有明显的优势。

因此可以得出结论,由于数据集分布的不同以及不平衡率的不同,不同隶属度函数结果存在差异,而从表中可以看出,文中提出的方法针对高不平衡率的数据具有很大的优势。

5　结束语

针对不平衡数据集,文中提出了一种改进的模糊支持向量机。在模糊隶属度构造过程中,根据各类样本点对于分类的贡献,将样本点分为四类:边界点、噪声点、孤立点、安全点。并使用紧密度区分出孤立点、噪声点,赋予较小隶属度值,再根据样本点之间的距离关系为安全点与边界点构造隶属度函数。该方法进一步降低了噪声点、孤立点对分类效果的影响。同时通过引入不平衡因子,有效提高了不平衡分类问题的分类精度。实验结果表明,该方法对于不平衡率较高的数据集具有良好的分类效果。

参考文献:

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

[2] 程　然. 最小二乘支持向量机的研究和应用[D]. 哈尔滨:哈尔滨工业大学,2013.

[3] Wu G,Chang E Y. Class-boundary alignment for imbalanced dataset learning[C]//Proc of 2003 workshop on learning from imbalanced data sets II. [s. l.]:[s. n.],2003:49-56.

[4] 赵相彬,梁永全,陈　雪. 基于支持向量机的不平衡数据分类研究[J]. 计算机与数字工程,2013,41(2):241-243.

[5] Lin Chunfu, Wang Shengde. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks,2002,13(2):464-471.

[6] 赵克楠,李　雷,邓　楠. 一种构造模糊隶属度的新方法[J]. 计算机技术与发展,2012,22(8):75-77.

[7] Batuwita R,Palade V. FSVM-CIL:fuzzy support vector machines for class imbalance learning[J]. IEEE Transactions on Fuzzy Systems,2010,18(3):558-571.

[8] Lakshmanan B,Priscilla A J,Ponni S,et al. Evaluation of im-

(下转第 48 页)

输出结构化字典。在 Extend YaleB 和 AR 人脸数据库上的实验表明,CA-KSVD 相对于其他方法,不仅在较小的字典尺寸下显著提高了识别性能,而且减少了字典冗余,得到了优化的字典原子。但是该算法学习得到的字典只具有良好的表示能力而不具备判别能力,因此在后续的研究中,可以考虑加入线性分类错误项或者判别型稀疏编码错误项等函数项,提高鉴别能力。

参考文献:

[1] 闫雪南,邹建成. 基于稀疏表示的人脸图像压缩方法[J]. 北方工业大学学报,2014,26(3):6-10.

[2] 雷 萌,张 环,王 弘. 基于哈希表的稀疏图像压缩算法研究[J]. 软件导刊,2013,12(9):50-52.

[3] 王良君,石光明,李 甫,等. 多稀疏空间下的压缩感知图像重构[J]. 西安电子科技大学学报,2013,40(3):73-80.

[4] 练秋生,陈书贞. 基于混合基稀疏图像表示的压缩传感图像重构[J]. 自动化学报,2010,36(3): 385-391.

[5] 练秋生,张 伟. 基于图像块分类稀疏表示的超分辨率重构算法[J]. 电子学报,2012,40(5):920-925.

[6] 吴一全,李 立,陶飞翔. 基于 Shearlet 域各向异性扩散和稀疏表示的图像去噪[J]. 应用科学学报,2014,32(3):221-228.

[7] 罗 晖,褚红亮,王世昌. 基于 K-SVD 的低信噪比 WMSN 视频图像稀疏去噪[J]. 计算机工程与科学,2014,36(3):497-501.

[8] 王国权,许国军,陈晓丹,等. 基于小波变换和稀疏表示的人脸识别方法研究[J]. 中国科技信息,2014(8):155-158.

[9] 朱 杰,杨万扣,唐振民. 基于字典学习的核稀疏表示人脸识别方法[J]. 模式识别与人工智能,2012,25(5):859-864.

[10] Wright J,Yang A Y,Ganesh A,et al. Robust face recognition via spare representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,31(2):210-227.

[11] Gao S H,Tsang L W H,Chia L T. Kernel sparse representation for image classification and face recognition[C]//Proc of 11th European conference on computer vision. Heraklion,Greece:[s. n.],2010:1-14.

[12] Zhang L,Yang M,Feng X C. Sparse representation or collaborative representation: which helps face recognition? [C]// Proc of IEEE international conference on computer vision. Barcelona:IEEE,2011:471-478.

[13] Aharon M,Elad M,Bruckstein A. K-SVD:an algorithm for designing overcomplete dictionaries for sparse representation [J]. IEEE Trans on Signal Processing,2006,54(11):4311-4322.

[14] Zhang Q,Li B X. Discriminative KSVD for dictionary learning in face recognition[C]//Proc of the IEEE conference on computer vision and pattern recognition. San Francisco, USA: IEEE,2010:2691-2698.

[15] Frigui H,Krishnapuram R. Clustering by competitive agglomeration[J]. Pattern Recognition,1997,30(7):1109-1119.

[16] Mallar S G,Zhang Z. Matching pursuits with time-frequency dictionaries[J]. IEEE Transactions on Signal Processing, 1993,41(12):3397-3415.

(上接第 43 页)

balanced datasets using fuzzy support vector machine-class imbalance learning(FSVM-CIL)[C]//Proc of international conference on recent trends in information technology. [s. l.]:[s. n.],2011:1131-1136.

[9] 边肇祺,张学工. 模式识别[M]. 第 2 版. 北京:清华大学出版社,2000.

[10] Veropoulos K,Campbell C,Cristianini N. Controlling the sensitivity of support vector machines[C]//Proc of the international joint conference on artificial intelligence. [s. l.]:[s. n.], 1999:55-60.

[11] Akbani R,Kwek S,Japkowicz N. Applying support vector machines to imbalanced datasets[C]//Proc of fifteenth European conference on machines learning. [s. l.]:[s. n.],2004: 39-50.

[12] 施化吉,周书勇,李星毅,等. 基于平均密度的孤立点检测研究[J]. 电子科技大学学报,2007,36(6):1286-1288.

[13] 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.

[14] Shao Y H,Chen W J,Zhang J J,et al. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification[J]. Pattern Recognition,2014,47(9):3158-3167.

面向不平衡数据的模糊支持向量机

作者： 刘凌， 郭剑， 韩崇， LIU Ling， GUO Jian， HAN Chong
作者单位： 南京邮电大学 计算机学院,江苏 南京,210003
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015,25(11)