





面到坐标轴原点的距离。

模糊支持向量机的决策函数为:

$$f(x) = \text{sign} \left\{ \sum_{i=1}^m \alpha_i^* y_i K(x, x_i) + b^* \right\} \quad (3)$$

$$0 \leq \alpha_i \leq \mu_i C \quad i = 1, \dots, l \quad (4)$$

$K(x, x_i)$ 为核函数, 常见的核函数有线性核函数、多项式核函数、高斯核函数等, 在求解过程中核函数的选择要视数据集性质而定。

利用不等距超平面距离改进后的模糊支持向量机。当  $0 < \lambda < 1$ , 超平面距离正类样本较近; 反之则超平面距离负类样本较近。改进后的模糊支持向量机最优决策面即为下列公式的最优解:

$$\min \frac{1}{1+\lambda} \|\omega\|^2 + C \sum_{i=1}^l u_i \varepsilon_i \quad (5)$$

不等式约束条件变为:

$$\begin{cases} y_i [(\omega^T x_i) + b] - \lambda + \varepsilon_i \geq 0 & y_i = 1 \\ y_i [(\omega^T x_i) + b] + 1 + \varepsilon_i \leq 0 & y_i = -1 \\ \varepsilon_i \geq 0 & i = 1, \dots, l \end{cases} \quad (6)$$

通过引入拉格朗日乘子求解上述不等式约束的凸优化问题:

$$\begin{aligned} L(\omega, b, \varepsilon, a, \beta) = & \frac{1}{1+\lambda} \|\omega\|^2 + C \sum_{i=1}^l u_i \varepsilon_i \\ & - \left\{ \sum_{y_i=1} a_i y_i [(\omega^T x_i + b) - \lambda + \varepsilon_i] \right. \\ & \left. + \sum_{y_i=-1} a_i y_i [(\omega^T x_i + b) - 1 + \varepsilon_i] \right\} - \sum_{i=1}^l \beta_i \varepsilon_i \end{aligned} \quad (7)$$

其中,  $a_i$ 为拉格朗日因子

求解的关键变为得到(7)的最小值, 故对(7)式中的  $\omega, b, \varepsilon$  分别求偏导得到:

$$\begin{cases} \frac{\partial L(\omega, b, \varepsilon, a, \beta)}{\partial \omega} = \frac{1}{2(1+\lambda)} \omega - \sum_{i=1}^l a_i x_i y_i = 0 \\ \frac{\partial L(\omega, b, \varepsilon, a, \beta)}{\partial b} = \sum_{i=1}^l a_i y_i = 0 \\ \frac{\partial L(\omega, b, \varepsilon, a, \beta)}{\partial \varepsilon_i} = u_i C - a_i - \beta = 0 \end{cases} \quad (8)$$

将式(8)中得到的结果代入到式(7), 利用拉格朗日对偶性可以将求解原问题满足约束条件的极小值转化为:

$$\max \sum_{y_i=1} \lambda a_i + \sum_{y_i=-1} a_i - \frac{1+\lambda}{4} \sum_{i=1}^l \sum_{j=1}^l y_i y_j a_i a_j (x_i \cdot x_j) \quad (9)$$

由  $\sum_{i=1}^l y_i a_i = 0$  得到于  $\sum_{j=1}^l a_j = \sum_{j=-1}^l a_j$ , 化简式(9)得到:

$$\begin{aligned} & \sum_{y_i=1} \lambda a_i + \sum_{y_i=-1} a_i - \frac{1+\lambda}{4} \sum_{i=1}^l \sum_{j=1}^l y_i y_j a_i a_j (x_i \cdot x_j) = \\ & \frac{1+\lambda}{2} \left\{ \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j a_i a_j (x_i \cdot x_j) \right\} \end{aligned} \quad (10)$$

将上面列出的凸优化问题求解完毕, 得到改进的模糊支持向量机的决策函数:

$$f(x) = \text{sign} \left\{ \frac{1+\lambda}{2} \left[ \sum_{x_i \in SV} a_i y_i K(x_i, x) + b^* \right] \right\} \quad (11)$$

$$0 \leq a_i \leq u_i C \quad i = 1, \dots, l \quad (12)$$

$\lambda$  的值影响超平面与类之间的空间距离, 若  $0 < \lambda < 1$ , 则超平面与正类间的空间距离较小; 若  $\lambda > 1$ , 则超平面与负类之间的空间距离较小; 若  $\lambda = 1$  该算法等同于标准的模糊支持向量机。

从式(11)、式(12)可以得到改进后的模糊支持向量机和标准的模糊支持向量机的基本原理相同的结论, 可以将标准模糊支持向量机的训练方法应用于改进的模糊支持向量机上。

### 3 确定隶属度函数

在超平面的确定过程中, 并不是所有的样本点都能起到决定性作用的, 样本贡献度就是度量求解超平面所需的样本点的性质。图1展示了样本空间的分布状态, 其中深色区域中的样本贡献度较大, 区域外的样本贡献度法较小, 在求解过程中更有被识别为噪声数据的可能性, 影响超平面位置的选取。另外, 大部分的支持向量样本位于阴影部分。本文提出一种确定隶属度函数的方法, 既考虑到了样本内的距离关系, 又考虑到了样本之间的相互关系。

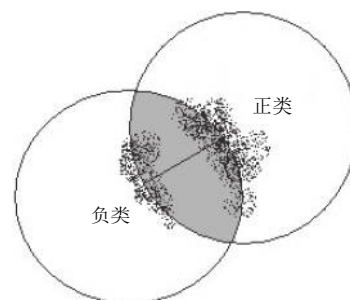


图1 样本的空间分布

通常定义隶属度函数如下:

定义 1. 类中心: 一类样本的平均样本特征定义为该类的中心. 如训练样本标记为:  $\{x_1, x_2, \dots, x_n\}$ , 类中心记为  $m = \frac{1}{n} \sum_{i=1}^n x_i$ , 正类样本的类中心记为  $m_+$ , 负类样本的类中心记为  $m_-$ .

定义 2. 两类样本之间的距离: 两类样本的类中心之间的距离为两类样本之间的距离, 记为  $d$ ,  $d = |m_+ - m_-|$ .

定义 3. 两类样本之间的互距离: 规定所有正类

样本到正类中心的距离  $d_{ip}^+ = |x_i - m_+|$ , 到负类中心的距离  $d_{ip}^- = |x_i - m_-|$ . 同样地, 规定所有负类样本到负类中心的距离  $d_{in}^- = |x_i - m_-|$ , 到正类中心的距离  $d_{in}^+ = |x_i - m_+|$ .

由于支持向量机是通过将样本映射到高维空间寻找最优决策面, 依据上文给出的定义, 各类的样本距离和样本互距离在高维空间中求解过程为:

$$m = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \tag{13}$$

$$d = |m_+ - m_-| = \left( \left| \sum_{x_+ \in X_+} \varphi(x_+) / n_+ - \sum_{x_- \in X_-} \varphi(x_-) / n_- \right|^2 \right)^{\frac{1}{2}}$$

$$= \left( \sum_{\substack{x_{+i} \in X_+ \\ x_{+j} \in X_+}} K(x_{+i}, x_{+j}) / n_+^2 + \sum_{\substack{x_{-p} \in X_- \\ x_{-q} \in X_-}} K(x_{-p}, x_{-q}) / n_-^2 - 2 \sum_{\substack{x_{+m} \in X_+ \\ x_{-n} \in X_-}} K(x_{+m}, x_{-n}) / n_+ n_- \right)^{\frac{1}{2}} \tag{14}$$

$$d_{ip}^+ = |\varphi(x_+) - m_+| = \left( \left| \varphi(x_+) - \sum_{x_+ \in X_+} \varphi(x_+) / n_+ \right|^2 \right)^{\frac{1}{2}} = \left( K(x_+, x_+) + \sum_{\substack{x_{+p} \in X_+ \\ x_{+q} \in X_+}} K(x_{+p}, x_{+q}) / n_+^2 - 2 \sum_{x_{+m} \in X_+} K(x_+, x_{+m}) / n_+ \right)^{\frac{1}{2}} \tag{15}$$

$$d_{ip}^- = |\varphi(x_+) - m_-| = \left( \left| \varphi(x_+) - \sum_{x_+ \in X_-} \varphi(x_-) / n_- \right|^2 \right)^{\frac{1}{2}} = \left( K(x_+, x_+) + \sum_{\substack{x_{-p} \in X_- \\ x_{-q} \in X_-}} K(x_{-p}, x_{-q}) / n_-^2 - 2 \sum_{x_{-m} \in X_-} K(x_+, x_{-m}) / n_- \right)^{\frac{1}{2}} \tag{16}$$

$$d_{in}^- = |\varphi(x_-) - m_-| = \left( \left| \varphi(x_-) - \sum_{x_- \in X_-} \varphi(x_-) / n_- \right|^2 \right)^{\frac{1}{2}} = \left( K(x_-, x_-) + \sum_{\substack{x_{-p} \in X_- \\ x_{-q} \in X_-}} K(x_{-p}, x_{-q}) / n_-^2 - 2 \sum_{x_{-m} \in X_-} K(x_-, x_{-m}) / n_- \right)^{\frac{1}{2}} \tag{17}$$

$$d_{in}^+ = |\varphi(x_-) - m_+| = \left( \left| \varphi(x_-) - \sum_{x_+ \in X_+} \varphi(x_+) / n_+ \right|^2 \right)^{\frac{1}{2}} = \left( K(x_-, x_-) + \sum_{\substack{x_{+p} \in X_+ \\ x_{+q} \in X_+}} K(x_{+p}, x_{+q}) / n_+^2 - 2 \sum_{x_{+m} \in X_+} K(x_-, x_{+m}) / n_+ \right)^{\frac{1}{2}} \tag{18}$$

为此提出了隶属度函数的设计算法如算法 1.

算法 1. 利用样本距离确定隶属度函数算法

- 1) 计算样本中心点之间的距离 $d$ ,计算正类样本的互距离 $d_{ip}^-$ ;
- 2) 比较样本距离与样本互距离的大小:若 $d_{ip}^- > d$ , 样本大都位于图 1 深色区域外部分,若 $d_{ip}^- \leq d$ , 样本大都位于图 1 深色区域内部分;
- 3) 取 $d_{ip}^- \leq d$ 的样本点计算其 $d_{ip}^+$ , 将其中的最大值记为 $R^+$ .
- 4) 同理得到负类样本的 $R^-$ .

最终得到两类样本的隶属度函数:

$$\begin{cases} s_{ip} = d_{ip}^+ / R^+, d_{ip}^- \leq d \\ s_{ip} = \delta, d_{ip}^- > d \end{cases} \quad y_i = 1 \quad (19)$$

$$\begin{cases} s_{in} = d_{in}^- / R^-, d_{in}^+ \leq d \\ s_{in} = \delta, d_{in}^+ > d \end{cases} \quad y_i = -1 \quad (20)$$

#### 4 实验结果与分析

当分类问题应用到现实生活中时,往往对负类的分类结果有更高的要求. 本文应用两种评价准则来验证改进算法的分类效果,即准确率和召回率. 其中准确率描述的是分类结果,表示负类分类结果中实际负类样本的比例. 召回率描述的是原有样本的分类覆盖率,表示的是原有样本中的负类被正确分类的比例. 其表达式分别为:

$$\text{准确率 Precision} = TN / (TN + FN)$$

$$\text{召回率 Recall} = TN / (TN + FP)$$

$TN$  代表实为负类且分类结果为负类的样本,  $FN$  代表实为负类但分类结果为正类的样本,  $FP$  代表实为正类但分类结果为负类的样本.

实验基于 UCI 数据集,并选出 4 种不平衡率不同的训练样本集,样本不平衡率如表 1 所示.

表 1 样本训练集

数据集	正类样本	负类样本	样本总数	不平衡比
Irist	2000	400	2400	5:1
Balance Scale	620	98	718	6.3:1
Yeast	13 100	1000	14 100	13.1:1
Abalone	6000	200	6200	30:1

实验将 IFD-FSVM 算法应用于 UCI 数据集验证算法性能,并将实验结果与 SVM、FSVM 在相同场景下的分类结果进行比较.

**SVM 算法:** 等距超平面且没有将隶属度函数应用于支持向量机.

**FSVM 算法:** 等距超平面线性隶属度函数的模糊支持向量机.

**IFD-FSVM 算法:** 应用不等距超平面距离的改进模糊支持向量机.

首先,将 4 种样本集的参数分别设置为 $\lambda_{Irist}=0.8$ ,  $\lambda_{Balance}=0.7$ ,  $\lambda_{Yeast}=0.63$ ,  $\lambda_{Abalone}=0.37$ ,  $\delta=0.2$  时,各样本集的分类结果如表 2 所示.

表 2 各数据集在 3 种算法下的分类准确率与回归率 (%)

数据集	SVM算法		FSVM算法		IFD-FSVM算法	
	准确率	回归率	准确率	回归率	准确率	回归率
Irist	74.34	90.01	82.73	89.95	90.02	93.89
Balance Scale	74.27	82.66	78.64	83.86	89.10	90.03
Yeast	62.75	84.37	75.05	89.92	82.06	95.16
Abalone	51.98	84.49	62.97	90.04	78.90	95.28

由表 2 可以看出,相比其他算法,IFD-FSVM 算法明显提高了分类准确率与回归率. 在 Irist 数据集上应用 IFD-FSVM 算法进行分类,准确率分别比应用标准 SVM 和 FSVM 提高了 15.68% 和 7.29%. 在 Balance Scale 数据集上应用本文算法进行分类,准确率分别比应用标准 SVM 和 FSVM 提高了 14.83% 和 10.46%. 而在 Yeast 数据集上,IFD-FSVM 算法的分类准确率比其他两种算法分别提高了 19.31% 和 7.01%,在 Abalone 数据集上则具有较大的准确度改善,较其他两种算法分别提高了 26.92% 和 15.93%.

各样本集的分类效果如图 2、图 3 所示.

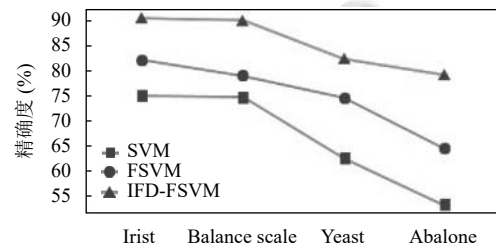


图 2 负类样本分类准确率对比

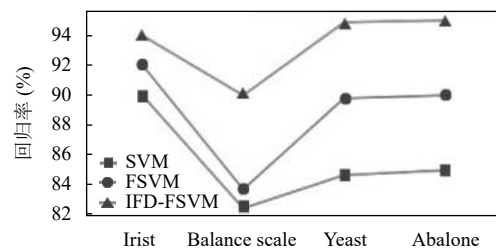


图 3 负类样本回归率对比

图 2 和图 3 分别展示了 3 种算法在 4 种样本集上负类的分类的准确率和召回率. 可以看出, IFD-FSVM 算法的分类效果明显优于另外两个标准算法. 且样本

数据不平衡比例越高,分类效果的改善越明显,在 Abalone 数据集上的负样本分类准确率和召回率都有较大幅度提升。

虽然参数 $\lambda$ 对分类器性能有着至关重要的影响,当参数 $\lambda < 1$ 时,负类的分类效果有明显改善,但并不是参数 $\lambda$ 的设置越小越好。当参数 $\lambda$ 过小时,正类的分类效果受到影响。如将4种样本集的参数分别设置为 $\lambda_{\text{Irist}}=0.21$ ,  $\lambda_{\text{Balance}}=0.19$ ,  $\lambda_{\text{Yeast}}=0.12$ ,  $\lambda_{\text{Abalone}}=0.08$ 时,各样本集的正类分类准确率如图4所示。

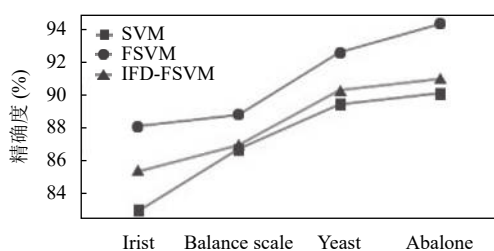


图4 正类样本分类准确率对比

从图4中可以看出,若将参数 $\lambda$ 的值设置为如上,相较于标准的模糊支持向量机,IFD-FSVM算法对正类分类效果明显下降。由于参数 $\lambda$ 过小,超平面与正类样本距离过小,负类样本被识别为噪声的概率增加,导致正类分类准确率受到影响。

## 5 结论与展望

通过对不平衡支持向量机的研究,本文提出了应用不等距超平面距离改进的模糊支持向量机 IFD-FSVM。算法通过改进原有的模糊支持向量机,引入参数 $\lambda$ 以调节超平面到正类的距离,实验时规定 $\lambda < 1$ ,令超平面接近正类样本。利用样本之间的互距离确定模糊隶属度函数,有利于确定贡献度大的样本数据,更好的反映了训练样本对超平面形成的贡献作用,降低了噪声数据给分类器性能带来的影响。最后利用UCI数据集来验证IFD-FSVM算法的有效性,实验结果说明IFD-FSVM算法能够有效提高不平衡样本的分类准确率。

## 参考文献

- Vapnik VN. Statistical Learning Theory. New York: Wiley, 1998.
- Bhalla VK, Kumar N. An efficient scheme for automatic web pages categorization using the support vector machine. New Review of Hypermedia and Multimedia, 2016, 22(3): 223–

242. [doi: 10.1080/13614568.2016.1152316]
- Zeng M, Zou BJ, Wei FR, et al. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS). Chongqing, China. 2016. 225–228. [doi: 10.1109/ICOACS.2016.7563084]
- 于巧, 姜淑娟, 张艳梅, 等. 分类不平衡对软件缺陷预测模型性能的影响研究. 计算机学报, 2018, 41(4): 809–824. [doi: 10.11897/SP.J.1016.2018.00809]
- Elleuch M, Maalej R, Kherallah M. A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition. Procedia Computer Science, 2016, 80: 1712–1723. [doi: 10.1016/j.procs.2016.05.512]
- 魏鑫, 张雪英, 李凤莲, 等. 面向非平衡数据集分类的改进模糊支持向量机. 计算机工程与设计, 2019, 40(11): 3124–3129.
- Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of 2005 International Conference on Advances in Intelligent Computing. Hefei, China. 2005. 878–887.
- 李勇, 刘战东, 张海军. 不平衡数据的集成分类算法综述. 计算机应用研究, 2014, 31(5): 1287–1291. [doi: 10.3969/j.issn.1001-3695.2014.05.002]
- 鞠哲, 曹隼喆, 顾宏. 用于不平衡数据分类的模糊支持向量机算法. 大连理工大学学报, 2016, 56(5): 525–531. [doi: 10.7511/dllgxb201605013]
- Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, USA. 2008. 213–220. [doi: 10.1145/1401890.1401920]
- Dal Pozzolo A, Caelen O, Johnson RA, et al. Calibrating probability with undersampling for unbalanced classification. Proceedings of 2015 IEEE Symposium Series on Computational Intelligence. Cape Town, South Africa. 2015. 159–166. [doi: 10.1109/SSCI.2015.3]
- 魏力, 张育平. 一种改进型的不平衡数据欠采样算法. 小型微型计算机系统, 2019, 40(5): 1094–1098. [doi: 10.3969/j.issn.1000-1220.2019.05.032]
- 张菲菲, 王黎明, 柴玉梅. 一种改进过采样的不平衡数据集成分类算法. 小型微型计算机系统, 2018, 39(10): 2162–2168. [doi: 10.3969/j.issn.1000-1220.2018.10.006]
- Li BY, Hu JL, Hirasawa K. Support vector machine classifier with WHM offset for unbalanced data. Journal of Advanced

- Computational Intelligence and Intelligent Informatics, 2008, 12(1): 94–101. [doi: 10.20965/jaciii.2008.p0094]
- 15 刘艳, 钟萍, 陈静, 等. 用于处理不平衡样本的改进近似支持向量机新算法. 计算机应用, 2014, 34(6): 1618–1621. [doi: 10.11772/j.issn.1001-9081.2014.06.1618]
- 16 Wang Z, Zhu YW, Chen ZZ, *et al.* Multi-view learning with fisher kernel and bi-bagging for imbalanced problem. Applied Intelligence, 2019, 49(8): 3109–3122. [doi: 10.1007/s10489-019-01428-1]
- 17 Malyscheff AM, Trafalis TB. Kernel classification using a linear programming approach. Annals of Mathematics and Artificial Intelligence, 2020, 88(1): 39–51.
- 18 Wang F, Liu SJ, Ni WC, *et al.* Imbalanced data classification algorithm with support vector machine kernel extensions. Evolutionary Intelligence, 2019, 12(3): 341–347. [doi: 10.1007/s12065-018-0182-0]
- 19 Li YJ, Guo HX, Liu X, *et al.* Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowledge-Based Systems, 2016, 94: 88–104. [doi: 10.1016/j.knosys.2015.11.013]
- 20 Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. Soft Computing, 2015, 19(12): 3369–3385. [doi: 10.1007/s00500-014-1291-z]
- 21 Lin Cf, Wang SD. Fuzzy support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 464–471. [doi: 10.1109/72.991432]
- 22 Zhang JJ, Zhong P. Learning biased SVM with weighted within-Class scatter for imbalanced classification. Neural Processing Letters, 2020, 51(1): 797–817. [doi: 10.1007/s11063-019-10096-8]
- 23 Liu KM, Wu XJ. Fuzzy support vector machines based on collaborative representation. Proceedings of the 11th International Conference on Natural Computation. Zhangjiajie, China. 2015. 64–68.
- 24 Zhou XL, Jiang PY, Wang XX. Recognition of control chart patterns using fuzzy SVM with a hybrid kernel function. Journal of Intelligent Manufacturing, 2018, 29(1): 51–67. [doi: 10.1007/s10845-015-1089-6]
- 25 杨志民, 王甜甜, 邵元海. 面向不均衡分类的隶属度加权模糊支持向量机. 计算机工程与应用, 2018, 54(2): 68–75. [doi: 10.3778/j.issn.1002-8331.1609-0112]