DP聚类的可信性加权模糊支持向量机

盛晓遐[1],杨志民[2],王甜甜[1]
1.浙江工业大学 理学院,杭州 310023
2.浙江工业大学 之江学院,杭州 310024

摘 要:由于SVM(Support Vector Machine)在有离群点和不平衡数据的问题中分类性能相对较低,有研究者提出了一种面向不均衡分类的隶属度加权模糊支持向量机,只是文中的模糊隶属度并不能较好衡量样本点对确定最佳分划超平面所做的贡献大小。针对以上问题提出了密度峰(Density Peaks,DP)聚类的可信性加权模糊支持向量机。首先由DP聚类找到离群点后剔除。再根据点到由DEC(Different Error Costs)确定的超平面的距离,得到初始隶属度,并用改进的FSVM-CIL(Fuzzy Support Vector Machines for Class Imbalance Learning)更新隶属度。之后剔除部分样本点,起到简约样本的作用,并减少数据不平衡带来的影响。通过实验验证了所提出算法的有效性。

关键词:离群点;不平衡数据;密度峰(DP);加权模糊支持向量机;模糊隶属度;可信性
文献标志码:A **中图分类号:**TP18;O159 doi:10.3778/j.issn.1002-8331.1804-0054

盛晓遐,杨志民,王甜甜.DP聚类的可信性加权模糊支持向量机.计算机工程与应用,2019,55(10):169-178.
SHENG Xiaoxia, YANG Zhimin, WANG Tiantian. DP clustering, creditability weighted fuzzy support vector machine. Computer Engineering and Applications, 2019, 55(10):169-178.

DP Clustering, Creditability Weighted Fuzzy Support Vector Machine

SHENG Xiaoxia[1], YANG Zhimin[2], WANG Tiantian[1]
1.College of Science, Zhejiang University of Technology, Hangzhou 310023, China
2.Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China

Abstract: Considering that SVM(Support Vector Machine) has relatively low classification performance in the case of outliers and unbalanced data, a weighted fuzzy support vector machine was proposed. And the fuzzy membership in that paper is not a good measure for the contribution of the sample to the determination of the optimal separating hyperplane. Thus, a DP(Density Peaks)clustering, creditability weighted fuzzy support vector machine is proposed. Outliers are found by DP clustering, then the outliers are eliminated. The distance from every sample to the hyperplane determined by DEC(Different Error Costs)is used to bulid the initial degree of membership. Then the degree of membership is updated with the improved FSVM-CIL(Fuzzy Support Vector Machines for Class Imbalance Learning). Finally, some samples are removed, which reduces the number of samples and reduces the impact of data imbalances. The effectiveness of the proposed algorithm is verified by experiments.

Key words: outliers; unbalanced data; Density Peaks(DP); weighted fuzzy support vector machine; fuzzy membership; creditability

1 引言

支持向量机(Support Vector Machine,SVM)是由Vapnik与其领导的贝尔实验室的研究小组于1995年开发出来的一种用于解决二分类问题的机器学习技术。它建立在VC维理论和结构风险最小化(Structural Risk Minimization,SRM)原则基础之上。在SVM中要求分

基金项目:国家自然科学基金(No.10926198);浙江省自然科学基金(No.LY16A010020)。
作者简介:盛晓遐(1992—),女,硕士研究生,研究方向为数据挖掘与支持向量机,E-mail:ctshengxiaoxia@126.com;杨志民(1957—),男,教授,博士生导师,研究方向为数据挖掘与支持向量机、不确定性信息处理等;王甜甜(1991—),女,硕士研究生,研究方向为数据挖掘与支持向量机。
收稿日期:2018-04-08 **修回日期:**2018-06-25 **文章编号:**1002-8331(2019)10-0169-10
CNKI **网络出版:**2019-01-17, http://kns.cnki.net/kcms/detail/11.2127.tp.20190115.1151.004.html

类间隔最大,实际上就是对推广能力的控制,是SVM的核心思想之一[1]。由于SVM出色的学习性能,该技术已经成为当前国际机器学习界的研究热点,并在较多领域得到应用,如语言识别[2]、金融预测[3]等。

然而,SVM对离群点敏感,以及它在处理不平衡数据集时由于分划超平面偏移而使得分类性能下降[4]。其中不平衡数据问题是当前研究的热点问题。

针对离群点问题,利用1965年Zadeh提出的模糊集理论[5],Lin等对二次规划的松弛变量添加模糊隶属度,为不同样本点构造相应的隶属度,从而构造模糊支持向量机模型,使得离群点的隶属度相应较小,从而减小它对确定分划超平面的影响。只是,如何选择一个合适的模糊隶属度模型也非易事[6]。利用1978年Zadeh提出的将模糊集作为可能性理论的基础想法[7],杨志民、邓乃扬于2007年提出基于可能性理论的模糊支持向量分类机[8],将训练样本点的输出转化为三角模糊数,并以可能性测度为基础,把模糊分类问题转化为模糊机会约束规划问题,较为充分地在数学本质上建立了模糊支持向量机模型。只是基于可能性理论的模糊支持向量分类机在处理模糊信息时有一定的缺陷,例如可能性测度为1的模糊事件未必一定成立,因此利用可能性测度建立模糊支持向量机,并且考虑极端情况(可能性测度为1)时容易出现误差[9]。

针对不平衡数据问题,研究者提出的方法主要有两种:一种是数据集的预处理,一种是算法改进。对于前者,研究者提出了欠采样[10]和过采样的方法。欠采样和过采样一定程度上可以缓解数据正负类样本点的数目不平衡问题,只是分别有可能会删除样本有用信息和增加样本额外信息。在过采样中,Chawla等于2002提出的SMOTE(Synthetic Minority Over-Sampling Technique)[11]是最知名的方法。还有对SMOTE改进的方法,如由Han等于2005年提出的Borderline-SMOTE,它与SMOTE不同的是,SMOTE对少数类样本都进行合成样本,而它只对靠近边界的点进行合成样本[12];另一种是由Rivera等人于2016年提出的新的SMOTE方法,即用正类点的支持向量并结合*K*NN(*K*-Nearest Neighbor)方法合成新的点[13]。这两种方法可以缓解原始SMOTE中合成对寻找超平面无贡献的点。另有由He等于2008年提出的ADASYN(Adaptive Synthetic Sampling),它根据不同的少数类的学习难易等级,确定出权重分布,从而合成样本,使得算法学习性能得到提高[14]。只是,以上这几种方法仍和SMOTE一样会增加样本额外信息。对于算法改进,1999年Veropolos提出了DEC(Different Error Costs),对正负类松弛变量给出不同惩罚因子的方法[15],从而有效提高算法准确率,只是该方法忽略了由于数据分布引起的不平衡问题;2006年,Imam等提出了zSVM,它是在决策函数的关于正支持向量的那一项前乘以权重z,使得分划超平面远离少数类样本点。文章致力于确定分划超平面的方向,从而使得在分划超平面和不同类之间保持一个良好的间隔,同时也保持分类性能[16]。只是该方法也忽略了由数据分布引起的不平衡问题。

针对SVM对离群点敏感以及在处理不平衡数据集时分类性能下降这两个问题,杨志民、王甜甜等提出了面向不均衡分类的隶属度加权模糊支持向量机[17](Weighted Fuzzy Support Vector Machine Faced on Fuzzy Membership of Imbalanced Classification,IFM-WFSVM)。利用样本点到以标准支持向量机(C-SVC)训练所得的分类超平面的距离构造隶属度,并更新隶属度,而后剔除远离超平面的样本点(即隶属度小于等于0.5的样本点)。起到了剔除离群点以及平衡正负类样本数目的作用。最后,将隶属度作用于可能性加权模糊支持向量机。该方法思想简单且易用算法实现,只是从文章的实验部分可以看出,IFM-WFSVM的效果并不是太好。主要原因是隶属度可能不能较好体现样本对最佳分划超平面的贡献程度。下面给出三点解释。

上面提到的IFM-WFSVM中最开始由标准支持向量机训练所得的分类超平面对最终隶属度的确定起到关键的作用。而这里的标准支持向量机(C-SVC)是在未剔除离群点且数据集不平衡的情况下确定的。它对离群点敏感以及在处理不平衡数据集时分类性能会下降。而且在更新隶属度的过程中,没有考虑到由于数据的正负类样本数目不均衡导致的分划超平面偏移。并且仅由点到超平面的距离确定隶属度,以此衡量样本点对最佳分划超平面的贡献程度不大全面。因此,最终得到的隶属度可能不能较好体现样本对最佳分划超平面的贡献程度。

本文针对IFM-WFSVM的不足,并综合考虑正负类样本数目、样本分布(点的密度、点所在类的分布及样本点到分划超平面的距离)等因素,以DP聚类、Yang等提出的模糊支持向量机(Fuzzy Support Vector Machine,FSVM)以及可信性理论为基础,提出了一种DP聚类的可信性加权模糊支持向量机(Density Peaks Clustering, Creditability Weighted Fuzzy Support Vector Machine,DP-CrWFSVM)。首先,DP聚类找到离群点,剔除离群点,消除离群点对算法的影响。其次,根据点到超平面(最初的超平面由DEC确定)的距离确定最初的隶属度。再将得到的隶属度作用于改进的FSVM-CIL(Fuzzy Support Vector Machines for Class Imbalanced Learning),得到分划超平面,根据点到超平面距离确定隶属度,如此更新隶属度。然后,对于少数类样本点,保留部分样本点的隶属度并调整剩余少数类样本点的隶属度。而对于多数类样本点,剔除部分样本点,并保留部分样本点的隶属度,不进行更新,同时根据样本点的分布更新剩余样本点的隶属度。其中剔除的样本点可看作是对找到最佳分划超平面无贡献的点,而且剔除部分样本点

一定程度上起到了简约样本以及修正因数据不平衡造成的分类误差的作用。最后,将得到的隶属度作用于可信性加权模糊支持向量机。

本文方法使用DP聚类找到离群点,很简便。而且根据点到超平面(最初的超平面由DEC确定)的距离确定最初的隶属度,相比较IFM-WFSVM而言,更加准确。因这里DEC是在去掉离群点后使用的,而且它在处理不平衡数据集时比标准支持向量机(C-SVC)有优势,而且后面还用到了改进的FSVM-CIL来更新隶属度,其中加入了平衡调节因子,从而减小了因数据不平衡带来的影响。然后利用样本分布(点的密度、点所在类的分布及样本点到分划超平面的距离)确定最终的隶属度,这样使得得到的隶属度相比较IFM-WFSVM而言能较好体现样本对最佳分划超平面的贡献程度。最后,在模糊环境下,可信性测度比可能性测度更能够准确表达模糊事件的状态[9]。最后,通过UCI数据集实验验证了本文算法在提升分类精度上的有效性。

2　面向不均衡分类的隶属度加权模糊支持向量机

2.1　基于可能性测度的模糊支持向量机

基于可能性测度模糊支持向量机的主要思想是对每个训练样本点赋予隶属度,用隶属度衡量样本对确定分划超平面所做的贡献的大小。引入三角模糊数,构造出新的训练集——模糊训练集,并以可能性测度为基础将分类问题模糊化。

给定输入空间的训练样本集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}(x_i\in\mathrm{R}^d)$,其中,$y_i\in\{-1,1\},i=1,2,\cdots,n$。这里使用一种特定的三角模糊数。定义正类样本点的隶属度为 $\delta_+(0.5\leqslant\delta_+\leqslant1)$,负类样本点的隶属度为 $\delta_-(0.5\leqslant\delta_-\leqslant1)$。

所对应的输出为三角模糊数[9]:

$$\tilde{y}=\begin{cases}\left(\dfrac{2\delta_+^{\,2}+\delta_+-2}{\delta_+},2\delta_+-1,\dfrac{2\delta_+^{\,2}-3\delta_++2}{\delta_+}\right),\\ \quad 0.5\leqslant\delta_+\leqslant1\\ \left(\dfrac{2\delta_-^{\,2}-3\delta_-+2}{-\delta_-},1-2\delta_-,\dfrac{2\delta_-^{\,2}+\delta_--2}{-\delta_-}\right),\\ \quad 0.5\leqslant\delta_-\leqslant1\end{cases}\tag{1}$$

对训练集做排序处理,得到如下的模糊训练集:

$$S=\{(x_1,\tilde{y}_1),(x_2,\tilde{y}_2),\cdots,(x_p,\tilde{y}_p),(x_{p+1},\tilde{y}_{p+1}),\cdots,(x_n,\tilde{y}_n)\}$$

其中,$(x_j,\tilde{y}_j),(j=1,2,\cdots,p)$ 为模糊正类点,而模糊负类点为 $(x_j,\tilde{y}_j),(j=p+1,p+2,\cdots,n)$。

引入适当的映射 $\phi:x_i\to\phi(x_i)$,将样本 x_i 映射到一个高维特征空间中。选取适当的核函数使得 $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j),i,j=1,2,\cdots,n$。引入松弛变量 ξ_i 以及惩罚项 C,这样对于某一置信水平 $\lambda(0<\lambda\leqslant1)$,在最小化经验风险的原则下,得到可能性模糊支持向量机的模型[9]:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i\\&\text{s.t. }Pos\{\tilde{y}_i((w\cdot\phi(x_i)+b)+\xi_i)\geqslant1\}\geqslant\lambda\\&\quad\xi_i\geqslant0,i=1,2,\cdots,n\end{aligned}$$

其中,$Pos\{A\}$ 为模糊事件 A 的可能性测度[9]。

该方法可以将对确定分划超平面的贡献大的样本点,即隶属度较大的样本点尽可能正确分类。并且忽略离群点,即忽略隶属度较小的样本点的影响,从而提高分类性能。但是,它在处理不平衡数据集的过程中可能会使得分划超平面偏移。

2.2　IFM-WFSVM模型

针对上述问题,杨志民、王甜甜等考虑将隶属度 δ_i 作为权重放在模糊支持向量机的目标函数中,并对模糊正类点和负类点赋予不同的惩罚参数 C_+、C_-,如此得到IFM-WFSVM模型,如下[17]:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\xi_i+C_-\sum_{j=p+1}^{n}\delta_j\xi_j\\&\text{s.t. }Pos\{\tilde{y}_i((w\cdot\phi(x_i)+b)+\xi_i)\geqslant1\}\geqslant\lambda\\&\quad\xi_i\geqslant0,i=1,2,\cdots,n\end{aligned}$$

其中,$\tilde{y}_i=(r_{i1},r_{i2},r_{i3}),\tilde{y}_j=(r_{j1},r_{j2},r_{j3})$,两者都是三角模糊数,如式(1)所示。在置信水平 $\lambda(0<\lambda\leqslant1)$ 下,上式的清晰等价规划如下:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\xi_i+C_-\sum_{j=p+1}^{n}\delta_j\xi_j\\&\text{s.t. }((1-\lambda)r_{i3}+\lambda r_{i2})(w\cdot\phi(x_i)+b)+\xi_i\geqslant1\\&\quad((1-\lambda)r_{j1}+\lambda r_{j2})(w\cdot\phi(x_j)+b)+\xi_j\geqslant1\\&\quad i=1,2,\cdots,p;j=p+1,p+2,\cdots,n;\xi_i\geqslant0\end{aligned}$$

构造拉格朗日函数,求得以上规划的对偶问题,得到最优解,最终得到决策函数。

2.3　模糊隶属度

可以看出模糊隶属度起到重要作用,如何得到模糊隶属度呢?这里采用的方法是首先利用标准支持向量机(C-SVC)进行训练,得到分划超平面后,计算点 x_i 到超平面的距离 d_i,从而得到隶属度,如下:

$$\delta_i=\begin{cases}1-\dfrac{d_i}{d_{\max}^{+}+\sigma},i=1,2,\cdots,p\\1-\dfrac{d_j}{d_{\max}^{-}+\sigma},j=p+1,p+2,\cdots,n\end{cases}\tag{2}$$

其中,$d_{\max}^{+}$、$d_{\max}^{-}$ 分别为正、负类样本点到超平面的最大距离;σ 为一任意小的正数。再将得到的隶属度作用于以下模型:

$$\begin{cases}\min\ \dfrac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\delta_i\xi_i\\\text{s.t. }y_i(w\cdot\phi(x_i)+b)+\xi_i\geqslant1\\\quad\xi_i\geqslant0,i=1,2,\cdots,n\end{cases}\tag{3}$$

这样重新得到分划超平面,并再次根据点到超平面

的距离得到隶属度,再作用于式(3),如此更新隶属度,直到相邻两次迭代所得结果相差小于或等于给定的误差阈值或者迭代次数达到设定的迭代次数阈值时,停止迭代。这里误差阈值取0.01,迭代次数阈值取5。如果最终隶属度小于或等于0.5,剔除样本点。这样做把对确定最佳分划超平面贡献小的点(包括离群点)剔除了,并起到了简约样本的作用。

2.4 平衡调节因子

综合考虑样本数量和分布情况,得到一种新的惩罚参数的计算方法[17]。样本数量可直接得出,而分布情况,这里用隶属度描述,因隶属度表现出点到超平面的距离大小。

要使得正负类样本的误差相对均衡,最好满足如下条件:

$$C_+^2\sum_{i=1}^{p}\delta_i^2=C_-^2\sum_{j=p+1}^{n}\delta_j^2$$

引入正负类样本数目 n^+、n^- 以及正负类样本点隶属度的平均值 $\bar{\delta}_+$、$\bar{\delta}_-$,得到:

$$C_+^2 n^+\bar{\delta}_+^2=C_-^2 n^-\bar{\delta}_-^2$$

因此可设惩罚参数:

$$C_+=C\bar{\delta}_-\sqrt{\frac{n^-}{n}},C_-=C\bar{\delta}_+\sqrt{\frac{n^+}{n}} \tag{4}$$

这里惩罚参数只用于IFM-WFSVM模型,并不用于式(3)。

2.5 IFM-WFSVM的优缺点

将最终的隶属度小于等于0.5的样本点剔除,起到了剔除离群点和简约样本的作用。同时,平衡调节因子的使用缓解了数据的不平衡性。只是,最开始由标准支持向量机(C-SVC)训练所得的分类超平面对最终隶属度的确定起到关键的作用。而这里的标准支持向量机是在未剔除离群点且数据集不平衡的情况下确定的。它对离群点敏感以及在处理不平衡数据集时分类性能会下降。而且在隶属度更新的过程中样本数目仍然不平衡,因平衡调节因子没有用于式(3),所以用式(3)确定的超平面可能也会偏斜。这里只使用样本点到超平面的距离衡量样本点的分布,可能只是单方面体现样本点分布。因此,最终得到的隶属度可能不能较好体现样本对最佳分划超平面的贡献程度。

3 DP聚类的可信性加权模糊支持向量机

3.1 DP聚类

与IFM-WFSVM不同的是,DP聚类的可信性加权模糊支持向量机(DP-CrWFSVM)在还未求最初超平面前就使用DP聚类找到离群点,然后剔除离群点,消除离群点对算法的影响。下面简单介绍DP聚类。

Alex和Alessandro于2014年在SCI上提出一种新的聚类方法——Clustering by Fast Search and Find of Density Peaks。这种聚类的核心思想是聚类中心比它的近邻有更高的密度,以及聚类中心与密度比它更高的点有较大的距离[18]。这种聚类算法简单易理解,有较好的聚类效果。但在样本量小于2 000的情况下,原文中提出的点的密度计算公式中 d_c 的变化会较大影响聚类效果,因此改进点的密度计算公式如下[19]:

$$\rho_i=\sum_{j\neq i}\exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{5}$$

其中,d_{ij} 是样本点 i 到样本点 j 的欧式距离;d_c 是截断距离,通过对所有 d_{ij} 进行升序排列,d_c 为该排列的1%~2%上的数,d_c 的选择使得各点近邻的平均数是样本点个数的1%~2%。

$$\mu_i=\min_{j:\rho_j>\rho_i}(d_{ij})$$

μ_i 表示的是密度比点 i 大的所有点中与点 i 的最小距离。以 ρ_i 为横坐标,以 μ_i 为纵坐标画决策图,ρ_i 小 μ_i 大的是异常点,两者都大的是聚类中心。归类时采用比自己密度大并且离自己最近的点的类标签一致。

3.2 DEC

用DP聚类剔除离群点后,根据点到超平面(最初的超平面由DEC确定)的距离确定最初的隶属度,如式(2)。在数据不平衡的情况下,使用DEC确定超平面要比标准支持向量机(C-SVC)更有优势。下面简单介绍DEC。

由Veropolos提出的DEC,是给正负类松弛变量添加不同的惩罚因子 C_+ 和 C_-。DEC的规划式如下:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\xi_i+C_-\sum_{j=p+1}^{n}\xi_j\\&\text{s.t. }y_i(w\cdot\phi(x_i)+b)+\xi_i\geqslant 1\\&\quad\xi_i\geqslant 0,i=1,2,\cdots,n\end{aligned}$$

其中正类点下标是 $1,2,\cdots,p$;负类点下标是 $p+1,p+2,\cdots,n$。

研究表明取 C_+/C_- 为数据集中负类样本数目与正类样本数目之比,将使得算法有效性提高。

3.3 改进的FSVM-CIL

将以上得到的隶属度再作用于改进的FSVM-CIL。下面简单介绍改进的FSVM-CIL。

Batuwita于2010年提出了FSVM-CIL[20]。它是FSVM(Fuzzy Support Vector Machines)与CIL(Class Imbalance Learning)的结合,分别使得SVM对离群点以及类不均衡问题较不敏感。本文对FSVM-CIL进行了改进,将IFM-WFSVM中使用的平衡调节因子加入其中,改进的FSVM-CIL的规划式如下:

$$\begin{cases}\min\ \dfrac{1}{2}\|w\|^2+C_+\displaystyle\sum_{i=1}^{p}\delta_i\xi_i+C_-\sum_{j=p+1}^{n}\delta_j\xi_j\\\text{s.t. }y_i(w\cdot\phi(x_i)+b)+\xi_i\geqslant 1\\\quad\xi_i\geqslant 0,i=1,2,\cdots,n\end{cases} \tag{6}$$

C_+、C_- 的求法如式(4)。

如此得到分划超平面后,根据点到超平面距离确定

隶属度,再作用于式(6),更新隶属度,直到相邻两次迭代所得结果相差小于给定的误差阈值或者迭代次数达到设定的迭代次数阈值时,停止迭代。这里误差阈值取0.01,迭代次数阈值取5。最后得到隶属度 δ_i''。平衡调节因子的加入可减小因数据不平衡带来的影响,这样的处理比IFM-WFSVM中隶属度的更新,如式(3),更加有优势。

3.4 最终隶属度的确定以及数据集的处理

得到新的隶属度 δ_i'' 后,对距离分划超平面足够近的点保留原来的隶属度,不进行更新,因这样的点对找到最佳分划超平面起重要作用。再根据样本点的分布(点的密度、点所在类的分布及点到分划超平面的距离)更新部分样本点的隶属度。这样相比较IFM-WFSVM中只使用点到分划超平面的距离确定隶属度要更加全面一些。并且删除部分多数类样本,也起到了改善样本不平衡比的作用。这里假设多数类是负类,最终隶属度的确定如下:

3.4.1 对于少数类即正类样本点

(1)当 $\delta_i''\geqslant\beta$ 时,说明样本点距离超平面足够近,保留样本点,隶属度不进行更新。

(2)其他情况时,根据点的密度即式(5)确定的隶属度:

$$\tau_{i\_\mathrm{pos}}=\frac{\rho_i}{\rho_{\max}}$$

其中,$\rho_{\max}$ 为所在类中的最大密度。

根据点所在类中的分布确定的隶属度:

$$\upsilon_{i\_\mathrm{pos}}=1-\frac{d(x_i,c_k)}{d_{\max}(x_i,c_k)+\sigma}$$

其中,$d(x_i,c_k)$ 为样本点到它所在类的中心的距离;$d_{\max}(x_i,c_k)$ 为类中最远的点到类中心的距离。

最终的隶属度为:

$$\delta^*_{i\_\mathrm{pos}}=\frac{\tau_{i\_\mathrm{pos}}+\upsilon_{i\_\mathrm{pos}}+\delta_i''}{3}$$

若 $\delta^*_{i\_\mathrm{pos}}\leqslant 0.5$,则将样本点的隶属度都调整为 ω,这里 $\omega>0.5$,这样调整是为了保留少数类样本点并使用模糊支持向量机的三角模糊数。

3.4.2 对于多数类即负类样本点

(1)当 $\delta_j''\geqslant\beta$ 时,说明样本点距离超平面足够近,保留样本点,隶属度不进行更新。

(2)当 $0.5<\delta_j''<\beta$ 时,根据上面少数类样本点求隶属度的方法,求得 $\delta^*_{j\_\mathrm{neg}}$。为了使隶属度可以用于模糊支持向量机,若 $\delta^*_{j\_\mathrm{neg}}\leqslant 0.5$,则取 $\delta^*_{j\_\mathrm{neg}}=\omega$(与3.4.1小节(2)中的 ω 值相同)。

(3)当 $\delta_j''\leqslant 0.5$ 时,删除样本点。

这样得到最终的隶属度。若多数类是正类,则按照3.4.2小节,更新多数类的样本隶属度,按照3.4.1小节更新少数类的样本隶属度。多数类中删除的样本点可以看作对找到最佳分划超平面无贡献的点。

3.5 可信性模糊支持向量机

可信性模糊支持向量机的主要思想是对每个训练样本点赋予隶属度,引入三角模糊数,构造出新的训练集——模糊训练集,与2.1节所述的不同点在于它是以可信性测度为基础将分类问题模糊化。

所对应的输出为三角模糊数,如式(1)所示,与2.1节一样,给训练集做相同的排序处理,引入适当的映射 $\phi:x_i\to\phi(x_i)$,将样本 x_i 映射到一个高维特征空间中。来选取适当的核函数从而使得 $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$,$i,j=1,2,\cdots,n$。引入松弛变量 ξ_i 以及惩罚项 C,这样对于某一置信水平 $\lambda(0.5<\lambda\leqslant 1)$,在最小化经验风险的原则下,得到可信性模糊支持向量机的模型[9]:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\delta_i\xi_i\\&\text{s.t. }Cr\{\tilde{y}_i((w\cdot\phi(x_i)+b)+\xi_i)\geqslant 1\}\geqslant\lambda\\&\qquad\xi_i\geqslant 0,i=1,2,\cdots,n\end{aligned}$$

其中,Cr 是可信性测度[9]。由于IFM-WFSVM中使用的基于可能性理论的模糊支持向量分类机在处理模糊信息时有一定的缺陷,例如可能性测度为1的模糊事件未必一定成立,因此利用可能性测度建立模糊支持向量机,并且考虑极端情况(可能性测度为1)时容易出现误差。而且在模糊环境下,可信性测度能够准确表达模糊事件的状态,因此选择基于可信性理论的模糊支持向量分类机。可信性模糊支持向量机是在支持向量机的基础上,根据训练样本的重要性为每个训练样本赋予不同的权重,即隶属度。隶属度越接近1表明样本越重要,这样使得越重要的样本分类正确,提高了分类性能。

3.6 DP聚类的可信性加权模糊支持向量机模型

为了减少因数据集不平衡引起的分类性能下降,为模糊正类点和模糊负类点赋予不同的惩罚参数 C_+、C_-,从而得到如下DP-CrWFSVM模型:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\xi_i+C_-\sum_{j=p+1}^{n}\delta_j\xi_j\\&\text{s.t. }Cr\{\tilde{y}_i((w\cdot\phi(x_i)+b)+\xi_i)\geqslant 1\}\geqslant\lambda\\&\qquad\xi_i\geqslant 0,i=1,2,\cdots,n\end{aligned}$$

在置信水平 $\lambda(0.5<\lambda\leqslant 1)$ 下,以上模型的清晰等价规划(即与其等价的普通规划)如下:

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C_+\sum_{i=1}^{p}\delta_i\xi_i+C_-\sum_{j=p+1}^{n}\delta_j\xi_j\\&\text{s.t. }(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})(w\cdot\phi(x_i)+b)+\xi_i\geqslant 1\\&\qquad(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})(w\cdot\phi(x_j)+b)+\xi_j\geqslant 1\\&\qquad i=1,2,\cdots,p;j=p+1,p+2,\cdots,n;\xi_i\geqslant 0\end{aligned}$$

这里平衡调节因子 C_+、C_- 的求法如式(4),只是其中使用的隶属度是最终得到的隶属度。

模糊正类点的输出为 $\tilde{y}_i=(r_{i1},r_{i2},r_{i3})$，模糊负类点的输出为 $\tilde{y}_j=(r_{j1},r_{j2},r_{j3})$。

该清晰等价规划是凸二次规划，构造拉格朗日函数，求得其对偶问题：

$$\min\ \frac{1}{2}(A+2B+D)-(\sum_{i=1}^{p}\alpha_i+\sum_{j=p+1}^{n}\beta_j)$$

$$\text{s.t.}\ \sum_{i=1}^{p}\alpha_i(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})+\sum_{j=p+1}^{n}\beta_j(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})=0$$

$$0\leqslant\alpha_i\leqslant C_-\delta_i,i=1,2,\cdots,p$$

$$0\leqslant\beta_i\leqslant C_-\delta_j,j=p+1,p+2,\cdots,n$$

其中：

$$A=\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})\times(2(1-\lambda)r_{j3}+(2\lambda-1)r_{j2})(\phi(x_i)\cdot\phi(x_j))$$

$$B=\sum_{i=1}^{p}\sum_{j=p+1}^{n}\alpha_i\beta_j(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})\times(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})(\phi(x_i)\cdot\phi(x_j))$$

$$D=\sum_{i=p+1}^{n}\sum_{j=p+1}^{n}\beta_i\beta_j(2(1-\lambda)r_{i1}+(2\lambda-1)r_{i2})\times(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})(\phi(x_i)\cdot\phi(x_j))$$

对偶问题的最优解为：

$$(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)^{\mathrm{T}}=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_p^*,\beta_{p+1}^*,\beta_{p+2}^*,\cdots,\beta_n^*)^{\mathrm{T}}$$

求得最佳分划超平面的法向量为：

$$\boldsymbol{w}^*=\sum_{i=1}^{p}\alpha_i^*(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})x_i+\sum_{j=p+1}^{n}\beta_j^*(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})x_j$$

若存在 β^* 的正分量使 $\beta_s^*\in(0,C_-\delta_s)$，则：

$$b^*=(2(1-\lambda)r_{s1}+(2\lambda-1)r_{s2})-(\sum_{i=1}^{p}\alpha_i^*(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})(\phi(x_i)\cdot\phi(x_s))+\sum_{j=p+1}^{n}\beta_j^*(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})(\phi(x_j)\cdot\phi(x_s)))$$

若存在 α^* 的正分量使 $\alpha_q^*\in(0,C_+\delta_q)$，则：

$$b^*=(2(1-\lambda)r_{q3}+(2\lambda-1)r_{q2})-(\sum_{i=1}^{p}\alpha_i^*(2(1-\lambda)r_{i3}+(2\lambda-1)r_{i2})(\phi(x_i)\cdot\phi(x_q))+\sum_{j=p+1}^{n}\beta_j^*(2(1-\lambda)r_{j1}+(2\lambda-1)r_{j2})(\phi(x_j)\cdot\phi(x_q)))$$

最终，得到决策函数为：

$$f(x)=\mathrm{sgn}((\boldsymbol{w}^*\cdot\phi(x))+b^*)$$

4 实验与结果分析

本文用实验结果验证了所提出算法的准确性和有效性。所有实验均在Intel® Core™ i5-6500 CPU 3.2 GHz 8 GB RAM PC机Matlab R2016a软件上实现。

为了验证使用DP方法找到离群点，剔除后再进行实验以及DP-CrWFSVM方法的有效性，本文采用7个取自UCI机器学习知识库的实际不平衡数据集进行实验，如表1所示。其中前4个数据集正类样本点少于负类样本点，后3个则相反。考虑到使用DP时需计算欧式距离，这里的数据集样本点属性都是数值型的（定量的）。将DP-CrWFSVM与10种算法分类性能以及训练时间进行对比，从而对DP-CrWFSVM算法的有效性进行说明。并且通过前后两种算法（SVC与NEW-SVC比较，C-SVC与NEW-C-SVC比较，DEC与NEW-DEC比较，依次类推）的比较，对先使用DP方法找到离群点，剔除后再进行实验的有效性进行了说明。

表1 实验所用7个数据集的情况

Dataset	Total	Dim	Pos	Neg	Ratio
Balances	625	4	49	576	11.755
Pima	768	8	268	500	1.866
liver	345	6	145	200	1.379
Musk(V1)	476	269	207	269	1.300
SpectF	267	44	212	55	0.259
WPBC	198	33	151	47	0.311
Haberman	306	3	225	81	0.360

本文采用 ACC、SE、SP 和 GM 来评价分类性能，其定义如下所示：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN},\ SE=\frac{TP}{TP+FN}$$

$$SP=\frac{TN}{TN+FP},\ GM=\sqrt{SE\times SP}$$

其中 TP、TN、FP、FN 分别为真正、真负、假正、假负的样本点数目。本文列出了核函数为非线性的实验结果。其中，非线性核函数 $K(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|^2}{p^2}\right)$ 为应用广泛的高斯径向基核函数。所有算法中的惩罚参数 C 和非线性核参数 p 的候选集为 $\{2^{-8},2^{-7},\cdots,2^8\}$，参数 β 与参数 ω 的候选集分别为{0.70,0.75,…,1.00}，{0.51,0.56,…,0.71}，参数 λ 的候选集为{0.50,0.60,…,1.00}，均取最优时的数值。若 λ 取0.50时最优，则再取 λ 的候选集为{0.51,0.52,…,0.55}进行实验，因在DP-CrWFSVM模型中 λ 取大于0.50的数。

以下是与本文提出的算法进行比较的算法。

SVC：不另外添加任何约束的标准支持向量分类机算法。

C-SVC：为松弛变量添加惩罚项。

DEC：为正负类松弛变量给出不同惩罚因子的方法。

FSVM：引入模糊事件理论，将训练样本点的输出转化为三角模糊数，并以可能性测度为基础，把模糊分

类问题转化为模糊机会约束规划问题,是数学意义上较为完全的模糊支持向量机模型。

FSVM-CIL:将FSVM与CIL结合的一种算法。

NEW-SVC:先用DP聚类的方法找到离群点,剔除后再用SVC进行实验。

NEW-C-SVC、NEW-DEC、NEW-FSVM与NEW-FSVM-CIL是在用DP聚类方法找到离群点并剔除离群点的基础上,再分别用C-SVC、DEC、FSVM、FSVM-CIL进行实验的算法。

4.1　算法准确率及稳定性的比较

本节具体给出了文章提出的DP-CrWFSVM算法与SVC、C-SVC、DEC、FSVM、FSVM-CIL及NEW-SVC、NEW-C-SVC、NEW-DEC、NEW-FSVM、NEW-FSVM-CIL算法的分类结果。以下列出了核函数为非线性时的实验结果,包括各算法对于各数据集的 ACC、SE、SP、GM 的均值和标准差。

从表2中可以看出,对于数据集liver,DP-CrWFSVM相比于表中列出的其他算法,它的 SE、GM 都较高(这里看重 SE 是因为liver数据集样本点数量正少负多,SE 较高表明正类点错分率低,而不平衡数据集分类期望少数类样本点错分率低)。对于数据集Haberman,DP-CrWFSVM相比于表中列出的其他算法,它的 SP、GM 都极高(这里看重 SP 是因为Haberman数据集样本点数量负少正多,SP 较高表明负类点错分率低,而不平衡数据集分类期望少数类样本点错分率低)。而在数据集Musk(V1)上,DP-CrWFSVM的 GM 第四高,并且和最高的只相差约0.28%,而 SE 也是第四高,和最高的相差约1.37%。只是在数据集WPBC上,DP-CrWFSVM的 GM 最低,并且和最高的相差约13.04%,而 SP 是第六高,和最高的相差约24.12%。

从表3中可以看出,对于数据集liver,除了算法NEW-FSVM,DP-CrWFSVM相比于表中列出的其他算法,它的 SE、GM 都较高。对于数据集Haberman,DP-CrWFSVM相比于表中列出的其他算法,它的 SP、GM 都极高。而在数据集Musk(V1)上,DP-CrWFSVM的 GM 第六高,并且和最高的只相差约1.26%,而 SE 也是第六高,和最高的相差约3.68%。在数据集WPBC上,DP-CrWFSVM的 GM 第六高,并且和最高的相差约14.54%,而 SP 是第六高,和最高的相差约23.82%。

同时,分析标准差可看出,只有在数据集liver上,

表2　DP-CrWFSVM算法与另5种算法实验结果对比表　%

Dataset	Index	SVC	C-SVC	DEC	FSVM	FSVM-CIL	DP-CrWFSVM
Balances	*ACC*	72.403 4±2.36	94.102 1±0.64	91.693 3±0.53	82.837 2±1.04	93.704 6±0.44	—
	SE	99.700 0±0.98	89.791 7±2.71	92.083 3±0.39	85.483 3±3.34	89.591 7±2.42	—
	SP	70.082 4±2.56	94.473 7±0.58	91.678 8±0.57	82.621 5±1.22	94.066 3±0.41	—
	GM	83.411 1±1.59	91.734 0±1.70	91.666 5±0.45	83.281 7±1.78	91.495 9±1.38	—
Pima	*ACC*	69.831 6±0.55	75.382 8±0.67	73.435 4±0.43	71.229 8±1.83	71.027 6±2.32	—
	SE	47.760 6±1.55	74.476 2±1.51	80.508 2±0.80	73.081 5±1.77	81.083 6±3.74	—
	SP	81.677 8±0.75	75.844 2±0.78	69.652 5±0.64	70.261 9±3.28	65.642 4±5.14	—
	GM	62.121 2±0.85	74.887 6±0.88	74.712 5±0.46	68.670 0±2.08	71.513 0±2.34	—
liver	*ACC*	60.315 9±1.31	69.517 7±1.29	69.638 0±1.27	68.764 2±1.17	68.856 6±1.29	73.905 7±1.06
	SE	56.855 2±2.88	72.288 5±2.15	70.276 3±2.10	56.836 7±2.16	63.128 9±2.44	74.628 0±1.58
	SP	63.149 0±1.97	67.618 9±2.12	69.198 8±2.19	77.487 2±2.09	73.170 0±2.80	73.327 7±1.97
	GM	59.053 0±1.81	69.094 8±1.58	69.016 1±1.42	65.644 2±1.18	66.764 6±1.41	**73.449 7±1.21**
Musk(V1)	*ACC*	95.859 2±0.31	95.975 3±0.37	96.007 5±0.43	94.007 5±1.02	94.039 2±0.75	95.832 6±0.60
	SE	95.067 7±0.48	95.284 2±0.55	95.333 9±0.72	89.643 1±2.44	89.387 1±1.87	93.960 2±1.02
	SP	96.435 6±0.46	96.534 9±0.40	96.547 7±0.48	97.422 9±0.56	97.717 3±0.60	97.432 6±0.65
	GM	95.699 2±0.32	95.871 2±0.41	95.897 7±0.46	93.261 9±1.32	93.264 0±1.03	95.619 7±0.65
SpectF	*ACC*	77.750 5±2.27	74.035 1±0.28	80.107 1±1.51	69.825 0±4.01	75.951 4±2.12	—
	SE	82.756 3±2.88	74.272 1±2.98	82.586 4±2.56	67.945 0±6.31	75.428 1±3.29	—
	SP	58.679 7±3.46	73.222 9±5.36	70.287 1±6.04	78.375 2±8.00	78.359 9±4.82	—
	GM	67.446 9±2.33	71.888 7±3.62	75.024 9±2.96	69.955 3±3.66	75.407 5±2.29	—
WPBC	*ACC*	70.853 8±2.61	74.728 0±2.55	75.923 5±2.09	75.253 1±2.57	66.161 2±3.97	75.975 3±1.92
	SE	72.295 0±3.70	79.092 4±3.68	79.781 1±3.21	80.182 7±3.09	65.981 8±6.00	86.926 2±2.53
	SP	64.299 0±6.44	62.148 7±6.14	64.764 2±4.56	61.209 1±4.13	68.747 6±6.83	44.624 2±4.89
	GM	65.738 9±4.60	66.679 9±4.53	69.535 7±3.65	67.408 8±3.79	61.931 0±4.62	56.497 4±5.81
Haberman	*ACC*	58.148 9±1.51	52.576 8±1.47	70.259 6±1.21	59.925 4±2.14	65.848 1±1.94	91.088 3±1.24
	SE	52.197 3±2.00	45.451 5±2.85	74.561 1±1.38	63.179 0±2.79	67.029 5±3.29	89.307 1±4.40
	SP	74.919 1±2.99	73.593 1±4.50	59.468 8±3.14	52.507 8±4.93	61.918 4±4.45	91.750 6±1.43
	GM	61.631 4±1.97	55.301 1±2.48	65.629 2±2.01	53.698 6±3.45	62.225 8±2.35	**89.236 5±3.28**

表3 采用DP聚类的各算法实验结果对比表 %

Dataset	Index	NEW-SVC	NEW-C-SVC	NEW-DEC	NEW-FSVM	NEW-FSVM-CIL	DP-CrWFSVM
Balances	*ACC*	76.921 3±2.01	93.230 3±0.97	90.052 5±0.73	76.831 6±0.94	93.441 3±0.61	—
	SE	96.475 0±4.55	89.075 0±2.61	90.433 3±1.02	86.833 3±4.15	84.950 0±4.26	—
	SP	75.472 1±2.14	93.545 1±1.02	89.970 7±0.78	76.047 8±1.12	94.118 7±0.60	—
	GM	84.162 3±4.13	90.861 8±1.76	89.854 5±0.78	80.392 5±2.54	88.638 4±2.97	—
Pima	*ACC*	77.051 3±0.76	79.694 5±0.37	80.995 3±0.30	74.568 6±2.33	79.565 8±1.22	—
	SE	67.422 3±1.32	82.417 0±0.61	78.477 5±1.07	77.930 9±2.00	74.492 8±2.63	—
	SP	83.239 0±0.86	77.943 7±0.60	82.617 0±0.78	72.375 2±3.82	82.838 3±2.51	—
	GM	74.627 5±0.86	80.008 1±0.40	80.339 0±0.42	70.693 9±4.33	77.895 2±1.37	—
liver	*ACC*	64.030 8±1.12	72.691 6±1.35	71.958 2±0.81	93.232 9±1.22	72.688 6±1.30	73.905 7±1.06
	SE	59.947 9±2.27	64.939 9±1.85	71.405 9±1.84	88.155 9±3.09	70.073 0±2.89	74.628 0±1.58
	SP	67.572 0±1.87	78.861 9±2.62	72.573 4±1.66	97.542 2±0.56	75.190 3±1.93	73.327 7±1.97
	GM	62.692 6±1.39	70.571 1±1.61	71.344 9±0.88	**92.523 0±1.54**	71.477 4±2.12	73.449 7±1.21
Musk(V1)	*ACC*	96.789 1±0.41	96.848 5±0.49	96.707 7±0.52	96.552 2±0.71	96.511 5±0.69	95.832 6±0.60
	SE	97.434 2±0.62	97.430 1±0.73	97.642 0±0.45	96.305 2±1.68	96.084 0±1.41	93.960 2±1.02
	SP	96.296 9±0.45	96.408 4±0.51	96.019 1±0.69	96.744 8±0.87	96.782 1±0.54	97.432 6±0.65
	GM	96.834 1±0.43	96.883 9±0.51	96.791 8±0.48	96.446 8±0.84	96.371 4±0.80	95.619 7±0.65
SpectF	*ACC*	75.252 6±1.68	70.911 7±2.45	78.475 1±1.54	74.681 9±2.21	71.753 7±2.23	—
	SE	75.274 6±2.39	67.671 7±3.50	78.244 1±1.76	75.354 6±3.19	68.292 9±3.11	—
	SP	74.918 7±4.29	85.517 6±3.29	79.964 7±4.33	71.527 2±6.24	86.577 6±5.71	—
	GM	72.967 1±3.69	74.228 6±4.15	77.446 5±2.89	70.581 0±4.23	75.170 6±2.91	—
WPBC	*ACC*	72.391 5±2.69	76.952 2±1.79	75.927 1±2.58	69.076 1±4.08	69.789 6±4.17	75.975 3±1.92
	SE	74.969 3±3.26	80.248 4±3.29	78.707 4±2.98	69.045 1±6.49	70.802 5±7.28	86.926 2±2.53
	SP	65.440 0±5.49	66.786 3±4.94	68.442 1±4.79	66.630 9±7.51	66.089 3±8.89	44.624 2±4.89
	GM	67.103 7±4.57	70.271 3±3.76	71.036 1±3.83	60.179 2±5.92	61.906 9±4.38	56.497 4±5.81
Haberman	*ACC*	67.785 0±2.30	77.334 4±1.83	79.327 7±1.55	71.553 7±1.68	77.741 8±0.80	91.088 3±1.24
	SE	64.264 4±3.36	87.096 6±2.81	88.260 0±1.92	73.614 6±2.65	84.084 6±1.45	89.307 1±4.40
	SP	76.369 4±2.86	54.688 4±3.60	58.658 0±3.58	67.099 8±4.42	62.620 8±2.78	91.750 6±1.43
	GM	68.431 3±2.81	66.108 4±3.48	70.531 4±2.68	68.732 0±2.58	71.086 3±2.12	**89.236 5±3.28**

表4 表3与表2比较结果表

Dataset	NEW-SVC		NEW-C-SVC		NEW-DEC		NEW-FSVM		NEW-FSVM-CIL	
	GM	*SE*;*SP*	*GM*	*SE*;*SP*	*GM*	*SE*;*SP*	*GM*	*SE*;*SP*	*GM*	*SE*;*SP*
Balances	+	–	–	–	–	–	–	+	–	–
Pima	+	+	+	+	+	–	+	+	+	–
liver	+	+	+	–	+	+	+	+	+	+
Musk(V1)	+	+	+	+	+	+	+	+	+	+
SpectF	+	+	+	+	+	+	+	–	–	+
WPBC	+	+	+	+	+	+	–	+	–	–
Haberman	+	+	+	–	+	–	+	+	+	+

DP-CrWFSVM有一定的优势。在另3个数据集Musk(V1)、WPBC及Haberman上,DP-CrWFSVM效果不佳。

表4是表3与表2比较的结果(SVC与NEW-SVC比较,C-SVC与NEW-C-SVC比较,DEC与NEW-DEC比较,依次类推)。表中左上角第一个加号表明,在数据集Balances上,NEW-SVC的*GM*比SVC高,反之减号表明NEW-SVC的*GM*比SVC低,依次类推,其他算法的比较可得相应加减号。在前4个数据集上(样本点数量都是正少负多),如果NEW-SVC的*SE*高于SVC,则是加号,否则是减号。在后3个数据集上(样本点数量都是负少正多),如果NEW-SVC的*SP*高于SVC,则是加号,否则是减号。

从表4中可以看出,整体上看加号多于减号,数量分别是70和18个,大约前者是后者4倍。在两种算法的比较中,后一种算法*GM*和*SE*或*SP*(前4个数据集用*SE*,后3个用*SP*)均高于前一种算法的情况占总数57%。因此,可知道新方法,即用DP聚类方法找到离群点,剔除后再做实验,效果较优。从表4中还可以得出,在数据集Balances上,后面的一种算法都不比前一种算法好。除了在数据集Balances,NEW-SVC在其余6个数据集上都比SVC效果好。而在数据集Musk(V1)上,后一种算法都比前一种算法效果好。

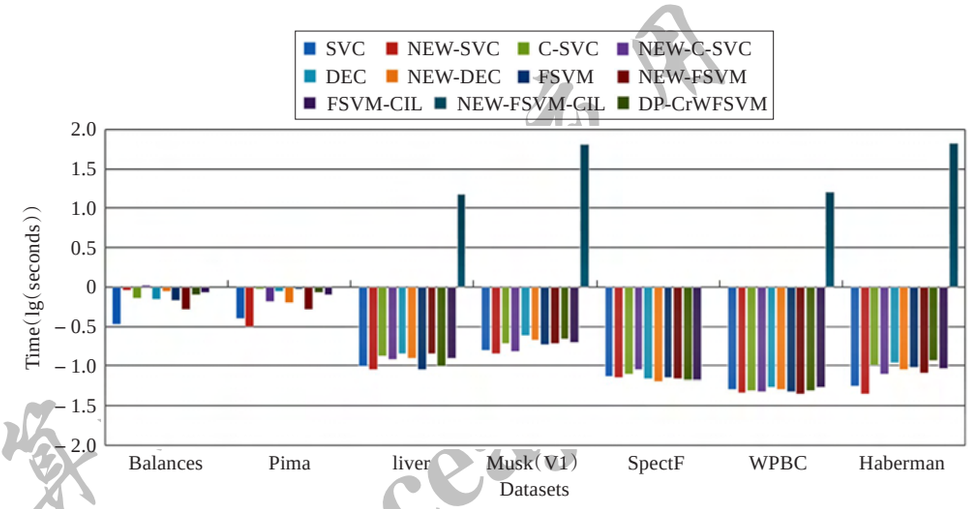

图1　各算法在7个数据集上的时间比较图

对结果进行综合分析可知,在算法稳定性上,DP-CrWFSVM效果相对不佳。在算法准确率上,用DP-CrWFSVM在某些数据集上得不到结果,在另一些数据集上效果不佳。然而,在数据集liver上,除了NEW-FSVM,与另外9种算法对比,DP-CrWFSVM都有较好结果。而且在数据集Haberman上,与另10种算法对比,DP-CrWFSVM有极好结果。这是因为所提出的方法不仅剔除了离群点,而且根据点到超平面(最初的超平面由DEC确定)的距离确定最初的隶属度。后面还用到了改进的FSVM-CIL,修正分划超平面,综合考虑了影响数据不平衡的因素,使得最终得到的隶属度能较好体现样本对最佳分划超平面的贡献程度。而且使用了可信性测度。可以得出用DP聚类的方法找到离群点,剔除后再做实验,效果较优(SVC与NEW-SVC比较,C-SVC与NEW-C-SVC比较,依次类推)。综合上述分析,DP-CrWFSVM使用DP方法找到离群点,剔除后再做实验,以及最初的超平面由DEC确定,并使用改进的FSVM-CIL更新隶属度,综合考虑正负类样本数目及分布的情况及使用可信性测度,使得该算法在某些数据集上得出更高的准确率。

4.2　算法时间的比较

本节列出了DP-CrWFSVM与其他10种算法在7个数据集中的训练时间对比(这里取lg(seconds))。NEW-SVC的训练时间是从DP聚类结束后开始记的,同理,其他几种剔除离群点的算法也是如此,除了DP-CrWFSVM。从图1可以看出,经过DP聚类剔除离群点的算法几乎都比未剔除离群点的训练时间要少(NEW-SVC与SVC比较,NEW-C-SVC与C-SVC比较,依次类推)。而DP-CrWFSVM在计算时间上高于其他所有算法。这是因为它使用了DP找出离群点,并且确定初始超平面,从而确定样本初始模糊隶属度,并不断更新超平面和隶属度,虽然剔除了部分样本,节约了时间,但前面消耗较多时间。

5　结束语

通过介绍IFM-WFSVM,指出问题,即它所得到的隶属度可能不会较好体现样本点对确定最佳分划超平面的贡献大小。并且给出三点原因:第一,最开始由标准支持向量机训练所得的分类超平面对最终隶属度的确定起到关键作用。而这里的标准支持向量机是在未剔除离群点且数据集不平衡的情况下确定的,它对离群点敏感以及在处理不平衡数据集时分类性能会下降。第二,在隶属度更新过程中样本数目仍然不平衡,而IFM-WFSVM中并没有对此问题予以重视。第三,使用样本点到超平面的距离衡量样本点的分布,可能只是单方面体现样本点分布。因此使得最终得到的隶属度不能较好体现样本点对确定最佳分划超平面的贡献大小。针对以上问题,提出了DP聚类的可信性加权模糊支持向量机(DP-CrWFSVM),使用DP聚类找到离群点后剔除,消除离群点对算法的影响。使用对不平衡数据集较少敏感的DEC代替标准支持向量机确定初始隶属度。将平衡调节因子应用于FSVM-CIL,使得在隶属度更新过程中减少因不平衡数据集带来的影响。将点的密度、点所在类的分布与点到分划超平面的距离结合来确定隶属度,使得隶属度体现样本较丰富的信息。如上所述,使得最终隶属度更能体现对确定最佳分划超平面的贡献大小。DP-CrWFSVM还剔除了多数类中部分对最佳分划超平面的确定贡献较小的样本,从而调节不平衡比,并且起到了简约样本的作用。DP-CrWFSVM最后是将最终的隶属度作用于可信性加权模糊支持向量机。在模糊环境下,可信性测度能够准确表达模糊事件的状态。通过实验验证了DP-CrWFSVM的有效性,在一些数据集上,采用某些算法前如果先通过DP聚类找出离群点并剔除离群点,再按照算法进行实验,使得与原来算法相比效果较优。未来研究中,将更多了解欠采样方法(如何判断样本点对确定分划超平面的贡献程度的基础理论知识)以及探究更合适的分类性能评价方法。

参考文献：

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

[2] Campbell W M,Campbell J P,Reynolds D A,et al.Support vector machines for speaker and language recognition[J].Computer Speech & Language,2006,20(2/3):210-229.

[3] Shin K S,Lee T S,Kim H J.An application of support vector machines in bankruptcy prediction model[J].Expert Systems with Applications,2005,28(1):127-135.

[4] Akbani R,Kwek S,Japkowicz N.Applying support vector machines to imbalanced datasets[C]//European Conference on Machine Learning,Pisa,Sep 20-24,2004:39-50.

[5] Zadeh L A.Fuzzy sets[J].Information and Control,1965,8(3):338-353.

[6] Lin C F,Wang S D.Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks,2002,13(2):464-471.

[7] Zadeh L A.Fuzzy sets as a basis for a theory of possibility[J].Fuzzy Sets & Systems,1978,1(1):3-28.

[8] 杨志民,邓乃扬.基于可能性理论的模糊支持向量分类机[J].模式识别与人工智能,2007,20(1):7-14.

[9] 杨志民,刘广利.不确定性支持向量机——算法及应用[M].北京:科学出版社,2012.

[10] Yang Z,Gao D.An active under-sampling approach for imbalanced data classification[C]//5th International Symposium on Computational Intelligence and Design,2012:270-273.

[11] Chawla N V,Bowyer K W,Hall L O,et al.SMOTE:synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research,2002,16(1):321-357.

[12] Han H,Wang W Y,Mao B H.Borderline-SMOTE:a new over-sampling method in imbalanced data sets learning[C]//Proc Int'l Conf Intelligent Computing,2005:878-887.

[13] Rivera W A,Xanthopoulos P.A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets[J].Expert Systems with Applications,2016,66:124-135.

[14] He H,Bai Y,Garcia E A,et al.ADASYN:adaptive synthetic sampling approach for imbalanced learning[C]//IEEE International Joint Conference on Neural Networks,2008:1322-1328.

[15] Veropoulos K,Campbell C,Cristianini N.Controlling the sensitivity of support vector machines[C]//International Joint Conference on Artificial Intelligence,1999:55-60.

[16] Imam T,Kai M T,Kamruzzaman J.z-SVM:an SVM for improved classification of imbalanced data[C]//Australasian Joint Conference on Artificial Intelligence.Berlin,Heidelberg:Springer,2006:264-273.

[17] 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[J].计算机工程与应用,2018,54(2):68-75.

[18] Alex R,Alessandro L.Clustering by fast search and find of density peaks[J].Science,2014,344:1492-1496.

[19] Xie J,Gao H,Xie W,et al.Robust clustering by detecting density peaks and assigning points based on fuzzy weighted *K*-nearest neighbors[J].Information Sciences,2016,354(C):19-40.

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

（上接第102页）

[10] Lee J,Lee J,Hong J.How to make efficient decoy files for ransomware detection?[C]//Proceedings of the International Conference on Research in Adaptive and Convergent Systems,2017:208-212.

[11] Kharraz A,Robertson W,Balzarotti D,et al.Cutting the gordian knot:a look under the hood of ransomware attacks[C]//LNCS 9148:Proceedings of the 12th International Conference on Detection of Intrusions and Malware,and Vulnerability Assessment,2015:3-24.

[12] Kirda E.UNVEIL:a large-scale,automated approach to detecting ransomware(keynote)[C]//IEEE International Conference on Software Analysis,Evolution and Reengineering,2017.

[13] Luo X,Liao Q.Awareness education as the key to ransomware prevention[J].Information Systems Security,2007,16(4):195-202.

[14] 王志海,童新海,沈寒辉.OpenSSL与网络信息安全:基础、结构和指令[M].北京:清华大学出版社,2007.

[15] Hunt G,Brubacher D.Detours:binary interception of Win32 functions[C]//Third USENIX Windows NT Symposium,1999.

[16] Guilfanov I.IDA fast library identification and recognition technology(FLIRT technology):in-depth[EB/OL].(2012-02-27)[2012-03-11].http://www.hex-rays.com/products/ida/tech/flirt/in-depth.shtml.