文章编号:0253-2778(2020)05-0596-09

非均衡数据情形的一种协同正则化多视图半监督学习分类器

崔文泉,陈 伟,程浩洋

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:利用多视图学习、流形学习以及协同正则化的多重惩罚处理,对含有大量无标签的类别数据提出一种多视图半监督学习的分类器构造方法.该方法由递归提升的方式对数据进行逐次多视图半监督学习,通过适当的标签化、均衡化处理改进每次集成的学习效率直到稳定.通过最小二乘和多分类SVM研究了新方法的性质,给出泛化误差的一个有意义上界,体现了新方法良好的泛化能力.模拟研究和实证分析显示,在有限样本情形下新方法具有良好的表现.

关键词:半监督学习;多视图学习;协同正则化;非均衡数据;集成学习

中图分类号:O212.1 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.05.007

2010 Mathematics Subject Classification:Primary 62H30;Secondary 6207

引用格式:崔文泉,陈伟,程浩洋. 非均衡数据情形的一种协同正则化多视图半监督学习分类器[J]. 中国科学技术大学学报,2020,50(5):596-604.

CUI Wenquan, CHEN Wei, CHENG Haoyang. A Multi-view based semi-supervised classifier with co-regularization for imbalanced data[J]. Journal of University of Science and Technology of China, 2020, 50(5):596-604.

A multi-view based semi-supervised classifier with co-regularization for imbalanced data

CUI Wenquan, CHEN Wei, CHENG Haoyang

(*Department of Statistics and Finance, School of Management, University of Science and of Technology of China, Hefei* 230026, *China*)

Abstract: A method of constructing a multi-view semi-supervised learning classifier was presented for manifold learning and multi-puncture processing. The multi-view and semi-supervised learning of the data is achieved through recursive optimization, and appropriate labeling and equalization processing, until the efficiency of learning becomes stable. The properties of this multi-classifier were given, for instance, an upper bound of the generalization error, which showed a good capacity for generalization. Simulation and empirical analysis showed that the new method performs well with small samples.

Key words: semi-supervised learning; multi-view learning; co-regularization; imbalanced data; ensemble learning

收稿日期:2019-04-14;**修回日期**:2019-05-17

基金项目:国家自然科学基金(71873128)资助.

作者简介:崔文泉(通讯作者),男,1964年生,博士/副教授.研究方向:数理统计.E-mail:wqcui@ustc.edu.cn

0　引言

分类是统计机器学习的一类重要方法，常见的分类模型如逻辑回归、决策树、支持向量机、随机森林等方法已有着广泛应用. 分类需要对样本做标记得到相应的标签，而标签的获取在某些场合下需要耗费一定的物力和人力，因此获取大量有标签样本相对比较困难，不过随着计算机技术的发展，大量无标签样本更易获得. 若只采用少量有标签数据，那么有监督学习训练得到的模型不具有很好的泛化能力，同时会浪费大量的无标签数据. 因此，研究综合利用少量有标签样本和大量无标签样本来提高分类性能的半监督学习(semi-supervised learning)成为当前机器学习研究的重要问题之一[1].

半监督学习有两个常用的基本假设，用以建立预测样本和学习目标之间的关系，即聚类假设、流形假设[2]. 根据学习方式的不同，半监督学习有基于协同训练、基于图的正则化以及基于 SVM 等方法. Blum 和 Mitchell 提出了协同训练方法[3]，对有标签数据和无标签数据进行迭代训练，以此扩充有标签样本集，但该方法计算复杂度较高. Sun 等[4]利用遗传算法结合协同训练的方式降低了算法优化过程的计算复杂度. 基于图正则化的方法则是利用了流形假设，将图论的相关理论引入半监督学习方法，巧妙地将分类问题转化为图学习问题[5]. 基于 SVM 的方法是在 SVM 目标函数中添加正则项来约束无标签样本，具有代表性的方法是 Belkin 等[6]提出的 Laplace-SVM 模型，该模型通过结合 Laplace 流形正则项和 SVM 的目标函数，将无标签数据的流形结构信息导入传统的 SVM 模型，通过有标签样本和无标签样本共同作用逼近流形空间，得到有标签数据和无标签数据的最大间隔. 随着对半监督学习问题的深入研究，许多学者也将其与深度学习[7-8]、主动学习[9-10]和多视图学习[6,11]等方法结合，对无标签数据做标记，从而提高分类性能.

多视图学习是结合多个视图或角度的信息进行模型构建与学习，其较单视图学习得到的特征更加全面，尤其当数据信息不完整时，多视图特征能够进行信息互补. 利用不同视图间数据信息的相互补充，多视图学习方法受数据的噪声影响较小，分类性能通常比单视图学习好. 多视图学习[12]较早的方法有典型相关分析[13]和协同训练算法[3]，之后不断有学者对这两种方法进行理论研究和技术拓展，并将其应用于其他学习问题. Belkin 等[6]在协同训练方法基础上提出协同正则化多视图半监督方法，通过对多个视图添加正则项的方式在每个视图中学习分类器. Sun[14]构造了多视图 Laplace 算子 SVM 模型，将流形正则和多视图正则方法融入 SVM 模型，把有监督学习推广到多视图半监督学习. Minh 等[15]提出的多视图学习采用了协同正则化方法，基于输入数据的不同视角(数据特征或数据形态)构建目标函数，对视图间函数和视图内函数分别做正则化约束，以此控制不同视图间函数的差异性和每个视图中函数的光滑性.

Minh 等[15]提出的半监督学习采用了流形正则化的方法，通过给无标签数据添加流形正则项来学习输入空间的几何结构. 流形学习[16]通过寻找高维数据分布的内在规律性，即找到高维空间中的低维流形，并求出相应的嵌入映射，以实现降维或数据可视化. 对于分类问题，认为数据本质上分布于一个低维流形子空间，因此可以通过无标签样本估计数据的几何边缘分布，并将此作为正则项协助学习平滑的流形子空间[6]. Brouard 等[17]和 Minh 和 Sindhwani[18]通过构造算子的方式将一维情形的流形正则方法推广到多维情形.

Micchlli 和 Pontil[19]和 Mroueh 等[20]分别提出了单视图、有监督情形下的最小二乘方法和多分类 SVM 方法，Minh 等[15]将这两种模型推广到半监督、多视图情形，并且给出了在平方损失和多分类 SVM 损失下的结果. 上述方法不是针对非均衡数据情形的，非均衡数据是一种非常重要的数据类型，其广泛存在于多个场景，如医学诊断领域的临床数据库中患病案例远少于正常案例[21]，诈骗案例(信用卡诈骗或手机诈骗)中合法用户比诈骗用户多[22]，其他诸如文本处理、风险管理等领域也存在数据非均衡现象. 本文将 Minh 等[15]提出的多视图半监督学习方法从数据均衡情形推广至数据非均衡情形.

针对数据非均衡问题，有通过欠抽样、过抽样方式对非均衡数据进行均衡化处理的基于抽样的方法，有根据错分类的代价矩阵对决策函数进行调整的基于代价敏感学习的方法[21]等. 本文采用基于抽样的方式将多数类数据随机分组，并利用每组数据和少数类样本训练子模型，最后对子模型进行集成. 在数据非均衡的半监督学习问题中，部分标签数据的缺失使得数据分组变得困难. 为了完成对无标签数据多数类样本的分组，本文提出一种利用有标签

数据及半监督学习的迭代方法,对无标签数据进行标签化处理,充分利用了无标签数据,有效解决了数据非均衡问题.

1 方法及性质

1.1 均衡化半监督集成学习方法(BSSEL 方法)

设观测数据 $z=\{(x_i,y_i)\}_{i=1}^{l}\cup\{x_i\}_{i=l+1}^{u+l}\in\mathcal{X}\times\mathrm{cl}(\mathcal{Y})\}$,$\mathrm{cl}(\mathcal{Y})=\{1,\cdots,P\}$,$P\geqslant 2$ 为类别个数.l 和 u 分别为有标签数据和无标签数据的数目.若问题为二分类情形,令 n_1,n_0 分别为样本中的正例数目和反例数目,当 $n_0\ll n_1$ 或 $n_0\gg n_1$ 时,称该问题是数据非均衡的.首先考虑多视图半监督模型[15],方法的实现转化为如下优化问题:

$$f_{z,\gamma}=\underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\frac{1}{l}\sum_{i=1}^{l}V(y_i,Cf(x_i))+\gamma_A\|f\|_{\mathcal{H}_K}^2+\gamma_B\langle f,M_Bf\rangle_{\mathcal{W}^{u+l}}+\gamma_W\langle f,M_Wf\rangle_{\mathcal{W}^{u+l}}\quad(1)$$

式(1)第一项刻画了有标签数据针对分类器 Cf 的经验损失,V 为凸损失函数.$f\in\mathcal{H}_K$ 刻画了多视图学习,$\mathcal{H}_K$ 为再生核 K 对应的再生核 Hilbert 空间.$C:\mathcal{W}\to\mathcal{Y}$ 为有界线性算子,它将多视图学习结果转化为分类的类别,即对 $\forall x\in\mathcal{X}$,$f_{z,\gamma}(x)\in\mathcal{W}$,有 $Cf_{z,\gamma}(x)\in\mathcal{Y}$.正则项中的 $f=(f(x_1),\cdots,f(x_{u+l}))\in\mathcal{W}^{u+l}$,$M_B$ 和 $M_W:\mathcal{W}^{u+l}\to\mathcal{W}^{u+l}$ 为对称正定算子,其对应的正则项 $\langle f,M_Bf\rangle_{\mathcal{W}^{u+l}}$ 和 $\langle f,M_Wf\rangle_{\mathcal{W}^{u+l}}$ 分别表示不同视图间函数的差异性和每个视图内函数的光滑性作惩罚约束,正则化参数 $\gamma_A>0,\gamma_B,\gamma_W\geqslant 0$.

本文中的损失函数 V 分别取平方损失和多分类 SVM 单纯形锥(simplex cone)[23]损失函数.当 V 取平方损失函数时,即

$$V(y_i,Cf(x_i))=\|r_{y_i}-Cf(x_i)\|_{\mathcal{Y}}^2,$$

则对 $\forall x\in\mathcal{X}$,预测 x 的类别为

$$\underset{1\leqslant k\leqslant P}{\operatorname{argmax}}\langle e_k,Cf_{z,\gamma}(x)\rangle_{\mathcal{Y}},$$

其中,$\mathcal{Y}=\mathbb{R}^P$.r 为一个映射:$\{1,\cdots,P\}\to\mathbb{R}^P$ 使得 $r_{y_i}^{\mathrm{T}}=(-1,\cdots,1,\cdots,-1)_{1\times P}$,$1\leqslant i\leqslant l$,若 x_i 属于第 k 类,则 r_{y_i} 第 k 个分量为 1,其余为 -1.$e_k^{\mathrm{T}}=(0,\cdots,1,\cdots,0)_{1\times P}$,第 k 个分量为 1,其余为 0.由文献[15]的定理 2,定理 4 和定理 10,可得对 $v\in\mathcal{X}$,

$$\hat{f}_{z,\gamma}(v)=\mathrm{vec}(RA^{\mathrm{T}}G[v,x])^{\mathrm{T}}\quad(2)$$

上式中,R 为对称半正定矩阵,核矩阵

$$G[v,x]=(G(v,x_i))_{1\leqslant i\leqslant l+u},$$

核函数 $G(\cdot,\cdot)$ 可用多项式核、高斯径向基核等,A 由下式求得

$$BAR+l\gamma_AA=Y_C\quad(3)$$

$$\mathrm{vec}(Y_C{}^{\mathrm{T}})=C^*r\quad(4)$$

$$B=((J_l^{u+l}\otimes cc^{\mathrm{T}})+\gamma_B(I_{u+l}\otimes M_m)+l\gamma_WL)G[x]\quad(5)$$

式中,$r=(r_{y_1},\cdots,r_{y_l})^{\mathrm{T}}$,$\otimes$ 表示 Kronecker 张量积.J_l^{u+l} 为 $(l+u)\times(l+u)$ 的对角阵,前 l 个对角元为 1,其余为 0.$c=(c_1,\cdots,c_m)^{\mathrm{T}}$ 为多视图的权重向量,m 为视图个数,C 取 $c^{\mathrm{T}}\otimes I_y$,$C^*=I_{(u+l)\times l}\otimes C^*$,$I_{(u+l)\times l}=[I_l,0_{l\times u}]^{\mathrm{T}}$,$I_l$ 为 $l\times l$ 的单位阵,C^* 为 C 的伴随算子.$M_m=mI_m-\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$,Gram 矩阵 $G[x]=(G(x_i,x_j))_{1\leqslant i,j\leqslant l+u}$.$L=(L_{i,j})_{1\leqslant i,j\leqslant l+u}$ 为分块矩阵,对角阵 $L_{i,j}=\mathrm{diag}(L_{ij}^1,\cdots,L_{ij}^m)$,其中 $L^i=D^i-A^i$,D^i 和 A^i 分别为第 i 个视图下权值矩阵 W^i 对应的无向图 T^i 的度矩阵和邻接矩阵.

当 V 是多分类 SVM 单纯形锥损失函数时,即

$$V(y_i,Cf(x_i))=\sum_{k=1,k\neq y_i}^{P}\max(0,-\langle s_k,s_{y_i}\rangle_{\mathcal{Y}}+\langle s_k,Cf(x_i)\rangle_{\mathcal{Y}}),$$

那么对 $\forall x\in\mathcal{X}$,预测 x 的类别为

$$\underset{1\leqslant k\leqslant P}{\operatorname{argmax}}\langle s_k,Cf_{z,\gamma}(x)\rangle_{\mathcal{Y}},$$

其中,$\mathcal{Y}=\mathbb{R}^{P-1}$,$s_k$ 为线性算子 S 中的向量.$S=[s_1,\cdots,s_P]$ 可取单纯形编码[15],单纯形编码为映射 $s:\{1,\cdots,P\}\to\mathbb{R}^{P-1}$ 使得① $\|s_k\|^2=1$;② $\langle s_j,s_k\rangle=-\frac{1}{P-1},j\neq k$;③ $\sum_{k=1}^{P}s_k=0$.由文献[15]的定理 2、定理 6 和定理 11,可得对 $v\in\mathcal{X}$,

$$\hat{f}_{z,\gamma}(v)=-\frac{1}{2}\mathrm{vec}(S\alpha^{\mathrm{opt}}(I_{(u+l)\times l}^{\mathrm{T}}\otimes c^{\mathrm{T}})M_{\mathrm{reg}}^{\mathrm{T}}G[v,x]^{\mathrm{T}})\quad(6)$$

式中,

$$\alpha^{\mathrm{opt}}=\arg\min_{\alpha\in\mathbf{R}^{P\times l}}\frac{1}{4}\mathrm{vec}(\alpha)^{\mathrm{T}}Q[x,C]\mathrm{vec}(\alpha)-\frac{1}{P-1}\mathbf{1}_{Pl}^{\mathrm{T}}\mathrm{vec}(\alpha),$$

$$Q[x,C]=(I_{(u+l)\times l}^{\mathrm{T}}\otimes c^{\mathrm{T}})G[x]\cdot M_{\mathrm{reg}}(I_{(u+l)\times l}\otimes c)\otimes S^*S,$$

$$M_{\mathrm{reg}}=[(\gamma_BI_{u+l}\otimes M_m+\gamma_WL)G[x]+\gamma_AI_{m(u+l)}]^{-1},$$

这里,S^* 为 S 的伴随算子.

本文将上述方法与集成思想相结合,将多视图半监督学习推广至非均衡情形.在文献[15]的多视图学习(multi-view learning,MVL)方法的基础上

提出一种均衡化半监督集成学习(balanced semi-supervised ensemble learning,BSSEL)方法,通过对无标签数据进行标签化处理,进而分组集成,降低模型预测的方差,提高模型的分类效果. BSSEL 方法利用训练样本中的有标签数据对无标签数据做标记,再将训练样本分成多组,以训练多个子分类器,然后对子分类器进行适当集成得到最终的分类器,基本步骤如下:

Step 1　设训练集 $D_z=D_l\cup D_u$,D_l 为有标签数据集,D_u 为无标签数据集. 用有监督学习方法对 D_l 进行分类器的构造,进而对 D_u 进行标签化处理,记标签化 D_u 后的数据集为 D_u^l ;

Step 2　将 D_l 中的多数类样本 D_{l_g} 随机分成 k 组:

$$D_{l_g}=\bigcup_{i=1}^{k} D_{g_i}^l\ ;$$

将 D_u^l 中的多数类样本 $D_{u_g}^l$ 随机分成 k 组:

$$D_{u_g}^l=\bigcup_{i=1}^{k} D_{u_{g_i}}^l\ ;$$

Step 3　将分组后的多数类样本与少数类样本组合:$D_{l_i}=D_{g_i}^l\cup D_{l_s}$,$i=1,\cdots,k$,$D_{l_s}$ 为 D_l 中的少数类样本;$D_{u_i}^l=D_{u_{g_i}}^l\cup D_{u_s}^l$,$i=1,\cdots,k$,$D_{u_s}^l$ 为 D_u^l 中的少数类样本;k 个训练子集为:$D_i=D_{l_i}\cup D_{u_i}^l$,$i=1,\cdots,k$;

Step 4　将 k 个子集 D_i 中 $D_{u_i}^l$($i=1,\cdots,k$)的标签信息去除,再将每组数据放入 MVL 模型中训练子分类器,并对无标签数据 D_u 进行预测,得到更新的 D_u^l ;

Step 5　重复 Step 2～4,直至预测结果稳定.

注:① Step 1 中有监督学习方法可以选取 SVM、随机森林、XGBoost、AdaBoost 等,本文采用的是文献[24]提出的"拟袋装"方法,该方法的实证结果显示在样本非均衡情况下,模型可以取得不错的分类结果,因此可以为无标签数据贴上较精准的初始标签.

② BSSEL 方法中视图的个数可以根据对变量作聚类的结果选取,本文对变量进行聚类的方法采用了 R 语言中 ClustOfVar 包的 K-means 聚类方法得到类别个数.

③ BSSEL 方法的惩罚参数 γ_A,γ_B,γ_W 及径向基核函数 σ 可以用 k 折交叉验证、留一交叉验证等方式选取,在样本量足够的情形下也可根据模型在验证集中的表现选择.

BSSEL 方法以文献[15]的多视图最小二乘学习(MVL-LS)和多视图 SVM 学习(MVL-SVM)为基分类器,根据损失函数的不同可分别记为 BSSEL-LS 方法和 BSSEL-SVM 方法,以下取径向基核作为再生核介绍 BSSEL-LS 分类方法,即算法 1.1. BSSEL-SVM 与 BSSEL-LS 类似,不再列出.

算法 1.1　BSSEL-LS 分类

训练数据 z;种类个数:P;

视图个数:m.

Parameters:

正则化参数 γ_A,γ_B,γ_W;

径向基核函数参数 σ;

权重向量 c.

Procedure:

Require:利用有标签数据及适当的有监督学习方法给无标签数据集 D_u 贴上初始标签 L^U;

令 diff=1,iter=1,$AUC_0=0$;

while diff>=0.001, iter <=50 do

(1) 分别将有标签数据和标签化后的数据中多数类进行分组,以 MVL-LS 为基分类器进行集成学习,对无标签数据预测,更新 L^U;

(2) 计算 AUC,diff=|AUC−AUC_0|,AUC_0=AUC;

(3) iter=iter+1;

end while

将有标签数据 D_l 和贴上标签的数据 D_u'分成 k 组:$\{D_{l_1},\cdots,D_{l_k}\}$,$\{D_{l_1}',\cdots,D_{l_k}'\}$;

for j=1:k do

求解基分类器的决策函数:

计算 $x_j=((x_i)_{i=1}^{u+l})_j$ 上的核矩阵 $G_j[x_j]$;

根据 $c^T\otimes I_P$ 计算矩阵 C;

根据 L 的定义[15]计算图 Laplace L_j;

根据式(4),(5)计算矩阵 Y_{C_j},B_j;

求解方程 $BAR+l\gamma_A A=Y_C$ 得到 A_j;

计算 v 和 x_j 的核矩阵 $G_j[v,x_j]$;

$\hat{f}_{j_{z,\gamma}(v)}=\mathrm{vec}(RA_j{}^TG_j[v,x_j]^T)\in\mathscr{Y}^m$;

计算决策函数 $h_{j_{z,\gamma}(v)}=C\,\hat{f}_{j_{z,\gamma}(v)}$;

end for

Ensure:将 k 组返回的结果进行集成,$h_{z,\gamma}(v)=\dfrac{1}{k}\sum_{j=1}^{k}h_{j_{z,\gamma}}(v)$.

注:① 算法中对无标签数据初始标签化的方法采用了文献[24]提出的"拟袋装"方法.

② BSSEL-SVM 与 BSSEL-LS 的不同之处在于,计算基分类器的决策函数时,前者使用式(6)得到决策函数 $h_{z,\gamma}(v)=S^TC\hat{f}_{z,\gamma}(v)$,最后采用同样

的集成方式得到结果.

③ BSSEL 算法的适用面较广,当 BSSEL 模型的损失函数 V 取逻辑回归、朴素贝叶斯等模型的损失函数时,只需更改求解基分类器决策函数的步骤即可,其他步骤不变.

④ BSSEL 方法充分利用已知信息,通过对子模型取均值的方法,降低模型预测的方差,提高预测精度,并且每组数据接近于均衡数据,能有效解决数据非均衡的问题.

1.2 BSSEL 方法的相关性质

本文进一步讨论了多视图最小二乘模型和多视图 SVM 模型的泛化误差,与文献[14,定理 3]类似,这里考虑损失函数分别为平方损失和多分类 SVM 单纯形锥损失的泛化误差.

定理 1.1 设

$$\begin{aligned}\mathscr{F}_1=\{&\tilde{f}:\tilde{f}(x,y)=\\&-\langle e_y,g(x)\rangle+\max_{y'\neq y}\langle e_{y'},g(x)\rangle,\\&g(x)=Cf(x)\in\mathscr{Y},f\in\mathscr{H}_K,\}\\\mathscr{F}_2=\{&\tilde{f}:\tilde{f}(x,y)=\\&-\langle s_y,g(x)\rangle+\max_{y'}\langle s_{y'},g(x)\rangle,\\&g(x)=Cf(x)\in\mathscr{Y},f\in\mathscr{H}_K\},\end{aligned}$$

其中,$y\in\{1,\cdots,P\}$,e_y,s_y 和 C 的定义见上文. 现从分布 $\mathscr{D}$ 中独立抽取 l 组观测数据 $D=\{(x_1,y_1),\cdots,(x_l,y_l)\}$,$y_i\in\{1,\cdots,P\}$,那么当损失函数为平方损失时,对于固定的 $\delta\in(0,1)$,以 $1-\delta$ 的概率,对任意 $g\in\mathscr{Y}$ 有

$$\begin{aligned}P_{\mathscr{D}}(y\neq\arg\max_{1\leqslant k\leqslant P}\langle e_k,g(x)\rangle)\leqslant&\\\frac{1}{l}\sum_{i=1}^{l}\|r_{y_i}-Cf(x_i)\|^2+&\\2\hat{R}_l(\mathscr{F}_1)+3\sqrt{\frac{\ln(2/\delta)}{2l}}&\end{aligned}\tag{7}$$

式中,$\hat{R}_l(\mathscr{F}_1)$ 是关于 $\mathscr{F}_1$ 的经验 Rademacher 复杂度.

当损失函数为多分类 SVM 单纯形锥损失时,对固定的 $\delta\in(0,1)$,以 $1-\delta$ 的概率,对任意 $g\in\mathscr{Y}$ 有

$$\begin{aligned}P_{\mathscr{D}}(y\neq\arg\max_{1\leqslant k\leqslant P}\langle s_k,g(x)\rangle)\leqslant\frac{1}{l}\sum_{i=1}^{l}\sum_{k=1,k\neq y_i}^{P}\xi_{k_i}^{+}+&\\2\hat{R}_l(\mathscr{F}_2)+3\sqrt{\frac{\ln(2/\delta)}{2l}}&\end{aligned}\tag{8}$$

式中,$\xi_{k_i}^{+}=(-\langle s_k,s_{y_i}\rangle+\langle s_k,Cf(x_i)\rangle)_+$,$\hat{R}_l(\mathscr{F}_2)$ 是关于 $\mathscr{F}_2$ 的经验 Rademacher 复杂度.

证明 首先考虑损失函数为平方损失的情形,令 $H(t)=\mathbf{1}\{t>0\}$,那么

$$P_{\mathscr{D}}(y\neq\arg\max_{1\leqslant k\leqslant P}\langle e_k,g(x)\rangle)=\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_1(x,y))],$$

其中,$\tilde{f}_1\in\mathscr{F}_1$,现考虑损失函数 $\mathscr{A}_1:\mathbb{R}\to[0,1]$,

$$\mathscr{A}_1(a)=\begin{cases}1,\ \text{if } a\geqslant 0;\\(1+a)^2,\ \text{if } -1\leqslant a\leqslant 0;\\0,\ \text{其他.}\end{cases}$$

由文献[14,引理 1]可得

$$\begin{aligned}&\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_1(x,y))-1]\leqslant\\&\mathbb{E}_{\mathscr{D}}[\mathscr{A}_1(\tilde{f}_1(x,y))-1]\leqslant\\&\hat{\mathbb{E}}[\mathscr{A}_1(\tilde{f}_1(x,y))-1]+\\&\hat{R}_l((\mathscr{A}_1-1)\circ\mathscr{F}_1)+3\sqrt{\frac{\ln(2/\delta)}{2l}},\end{aligned}$$

其中,$\hat{\mathbb{E}}[\cdot]$ 为样本的平均经验误差,$(\mathscr{A}-1)\circ\mathscr{F}=\{\mathscr{A}(\tilde{f}(x,y))-1:\tilde{f}(x,y)\in\mathscr{F}\}$. 则

$$\begin{aligned}\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_1(x,y))]\leqslant\hat{\mathbb{E}}[\mathscr{A}_1(\tilde{f}_1(x,y))]+&\\\hat{R}_l((\mathscr{A}_1-1)\circ\mathscr{F}_1)+3\sqrt{\frac{\ln(2/\delta)}{2l}}.&\end{aligned}$$

而

$$\begin{aligned}&\hat{\mathbb{E}}[\mathscr{A}_1(\tilde{f}_1(x,y))]\leqslant\\&\frac{1}{l}\sum_{i=1}^{l}(1-\langle e_{y_i},g(x_i)\rangle+\max_{y'\neq y}\langle e_{y'},g(x_i)\rangle)^2\leqslant\\&\frac{1}{l}\sum_{i=1}^{l}((1-\langle e_{y_i},g(x_i)\rangle)^2+\\&\sum_{k=1,k\neq y_i}^{P}(1+\langle e_k,g(x_i)\rangle)^2)=\\&\frac{1}{l}\sum_{i=1}^{l}\|r_{y_i}-g(x_i)\|^2=\\&\frac{1}{l}\sum_{i=1}^{l}\|r_{y_i}-Cf(x_i)\|^2,\end{aligned}$$

因为 $(\mathscr{A}_1-1)(0)=0$,那么由 $(\mathscr{A}_1-1)$ 的 Lipschitz 条件[25]有

$$\hat{R}_l((\mathscr{A}_1-1)\circ\mathscr{F}_1)\leqslant 2\hat{R}_l(\mathscr{F}_1).$$

综上可知式(7)成立.

接下来考虑损失函数为多分类 SVM 单纯形锥损失情形,与最小二乘情形类似,有

$$P_{\mathscr{D}}(y\neq\arg\max_{1\leqslant k\leqslant P}\langle s_k,g(x)\rangle)=\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_2(x,y))],$$

其中，$\tilde{f}_2 \in \mathscr{F}_2$，考虑损失函数 $\mathscr{A}_2:\mathbb{R}\to[0,1]$，

$$\mathscr{A}_2(a)=\begin{cases}1, \text{ if } a \geqslant 0;\\ 1+a, \text{ if } -1 \leqslant a \leqslant 0;\\ 0, \text{ 其他.}\end{cases}$$

同样的，由文献[14，引理 1]得

$$\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_2(x,y))-1] \leqslant$$
$$\mathbb{E}_{\mathscr{D}}[\mathscr{A}_2(\tilde{f}_2(x,y))-1] \leqslant$$
$$\hat{\mathbb{E}}[\mathscr{A}_2(\tilde{f}_2(x,y))-1]+$$
$$\hat{R}_l((\mathscr{A}_2-1)\circ\mathscr{F}_2)+3\sqrt{\frac{\ln(2/\delta)}{2l}},$$

那么

$$\mathbb{E}_{\mathscr{D}}[H(\tilde{f}_2(x,y))] \leqslant \hat{\mathbb{E}}[\mathscr{A}_2(\tilde{f}_2(x,y))]+$$
$$\hat{R}_l((\mathscr{A}_2-1)\circ\mathscr{F}_2)+3\sqrt{\frac{\ln(2/\delta)}{2l}},$$

而

$$\hat{\mathbb{E}}[\mathscr{A}_2(\tilde{f}_2(x,y))] \leqslant$$
$$\frac{1}{l}\sum_{i=1}^{l}(1-\langle s_{y_i},g(x_i)\rangle+\max_{y'}\langle s_{y'},g(x_i)\rangle)_+ \leqslant$$
$$\frac{1}{l}\sum_{i=1}^{l}(1+\sum_{k=1,k\neq y_i}^{P}\langle s_k,g(x_i)\rangle)_+ \leqslant$$
$$\frac{1}{l}\sum_{i=1}^{l}(\sum_{k=1,k\neq y_i}^{P}(\frac{1}{P-1}+\langle s_k,g(x_i)\rangle))_+ \leqslant$$
$$\frac{1}{l}\sum_{i=1}^{l}\sum_{k=1,k\neq y_i}^{P}(\frac{1}{P-1}+\langle s_k,g(x_i)\rangle)_+,$$

由于 $\langle s_k,s_{y_i}\rangle=-\frac{1}{P-1}(k\neq y_i)$，那么

$$\hat{\mathbb{E}}[\mathscr{A}_2(\tilde{f}_2(x,y))] \leqslant$$
$$\frac{1}{l}\sum_{i=1}^{l}\sum_{k=1,k\neq y_i}^{P}(-\langle s_k,s_{y_i}\rangle+\langle s_k,Cf(x_i)\rangle)_+ =$$
$$\frac{1}{l}\sum_{i=1}^{l}\sum_{k=1,k\neq y_i}^{P}\xi_{k_i}^{+}.$$

与最小二乘情形类似，有

$$\hat{R}_l((\mathscr{A}_2-1)\circ\mathscr{F}_2) \leqslant 2\hat{R}_l(\mathscr{F}_2).$$

综上可知式(8)成立. 证毕.

2　模拟实验与实证分析

以某银行信用卡中心提供的信用卡用户行为数据进行模拟实验和实证分析.

信用卡数据采用行为评分数据，有 148077 个用户. 其中违约客户(“坏”客户，正类)数量为 4605，非违约客户(“好”客户，负类)数量为 144012，“好”“坏”客户样本比为 34∶1. 数据的输入变量共有 298 个，将样本划分为训练集、测试集和验证集，因此采用 AUC 值衡量预测效果.

根据特征属性将变量划分成多个视图，利用 ClustOfVar 包对变量进行聚类，聚类结果为 4 类，每一类即为一个视图，每个视图的变量个数分别为 71，113，90，24，分别可视为客户在银行的办卡信息、客户的家庭信息、客户的工作信息和客户的其他信息.

2.1　模拟实验

实验每次随机抽取部分有标签数据，比较增加无标签数据后 BSSEL 方法的 AUC 值，并与其他有监督方法如 LDA、Naive Bayes、逻辑回归和 SVM 对比，其中调整的参数均根据模型在验证集上的表现选优.

从容量为 148077 的样本中取出 210 个个体作为有标签数据，分别取出容量为 105、420、1050、2100、3150 作为无标签数据，构成标签数据少、无标签数据多的 5 个数据情形：210＋105，210＋420，210＋1050，210＋2100 以及 210＋3150. 模拟研究是对此 5 个数据集分别进行的. 有标签的 210 个样本中：多数类容量为从 144012 个“好”客户中随机取出 200 个，少数类是从 4605 个“坏”客户中随机取出 10 个. 无标签数据是从上述剩余的 148077－210 个个体中获得的. 容量为 105 的无标签数据的获得：从 144012－200 个“好”客户中随机选取 100 个，以及从 4605－10 个“坏”客户中随机选取 5 个，将标签去掉得到容量为 105 的无标签数据；容量分别为 420、1050、2100、3150 的无标签数据的获得与此类似，其中(“好”，“坏”)客户分别为(400，20)、(1000，50)、(2000，100)、(3000，150).

将有标签数据的多数类样本随机划分成 10 组(每组 20 个)、20 组(每组 10 个)，并在 200 个样本中随机抽取 180 个数据，将其分为 30 组(每组 6 个)，这样不同情形下基分类器的正负样本比为 1∶2，1∶1，2∶1(近似). 从图 1 可以看出当基分类器的正负样本容量的比例为 1∶1 时 BSSEL 方法表现最好，因此以下结果分析均在此情形中讨论. 当正负样本容量比为 1∶2、2∶1(近似)时，本文同样对不同方法的效果加以分析对比，结果表明 BSSEL 方法随着无标签数据的增加表现相对最好，限于篇幅，这里不再给出.

从图 2 可以看出，传统的有监督方法无法利用

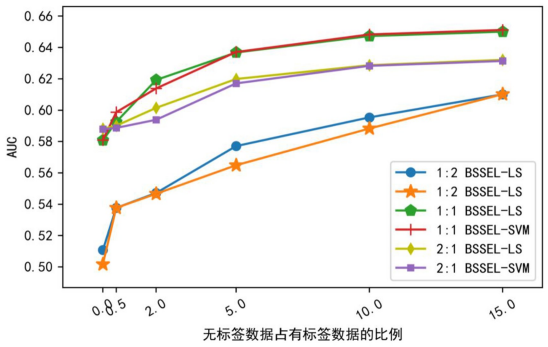


图 1 无标签数据增加时 BSSEL-LS 和 BSSEL-SVM 的 AUC 值

Fig. 1 AUC values of BSSEL-LS and BSSEL-SVM when unlabeled data increases

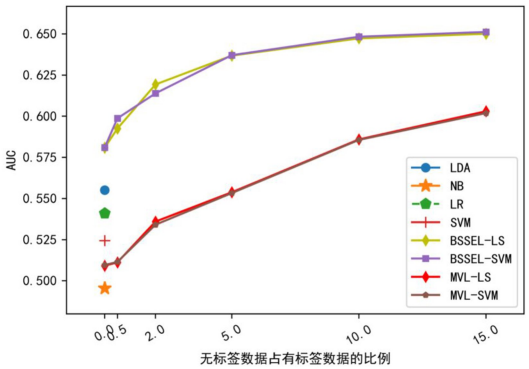
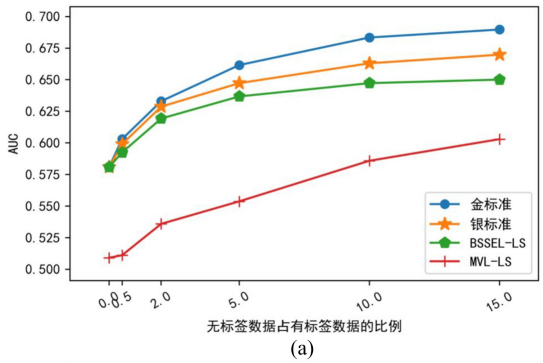


图 2 无标签数据增加对 AUC 值的影响

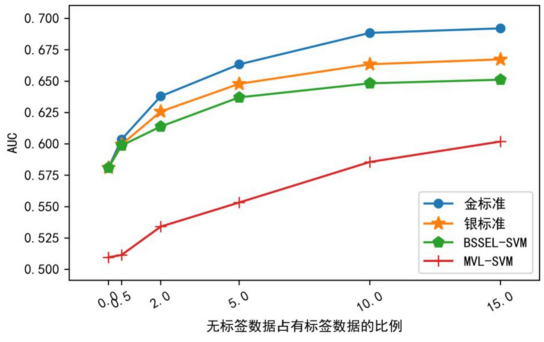
Fig. 2 Effect on AUC values when unlabeled data increases

无标签数据信息，只能训练有标签样本，因此 AUC 值保持不变. 当我们对无标签数据标签化、均衡化后，可以在半监督框架下分组集成，因此 BSSEL 方法在增加大量无标签数据时有良好的表现. 同时可以看出文献[15]的多视图方法 MVL 在处理非均衡数据时表现比 BSSEL 方法差.

我们将无标签数据的原有标签信息利用上，对全部数据进行有监督学习，学习的结果称为“金标准”；只在随机分组时利用无标签数据的原有标签信息，如此进行半监督学习的结果称为“银标准”. 显然“金标准”的结果最好，“银标准”的方法次之，图 3 是本文方法与“金标准”、“银标准”的模拟比较. 从图 3 可见，有监督集成学习的效果(“金标准”)最好，该方法利用了全部的有标签信息，且采用了集成学习方式，其次为半监督集成学习的效果(“银标准”)，BSSEL 方法对无标签数据标签化和均衡化后得到的分类效果和“银标准”相差不大. 若不考虑集成方式，直接将有标签数据和无标签数据放入 MVL 模型中，其表现最差.



(a)



(b)

图 3 基分类器多数类与少数类样本容量比 1∶1，“金标准”、“银标准”、BSSEL、MVL 的 AUC 值

Fig. 3 AUC values of methods ‘gold standard’, ‘silver standard’, BSSEL and MVL when sample size ratio of majority class to minority class is 1∶1 for base classifier

表 1 给出 BSSEL-LS 方法在不同惩罚参数下的效果，采用的数据与前文所述的模拟数据相同. $\gamma_B = \gamma_W = 0$ 时，无法利用无标签数据，只能使用有标签数据进行有监督的学习；当 $\gamma_B = 0, \gamma_W \neq 0$ 时，只对单个视图进行正则化处理；$\gamma_W = 0$，$\gamma_B \neq 0$ 时，只对多个视图之间的关系进行正则化处理；当 γ_B 和 γ_W 均不为 0 时，即为本文的多视图半监督学习方法. 其中，非零的调整参数均利用验证集进行优选.

表 1 模拟中 BSSEL-LS 在多重惩罚参数下的 AUC 值

Tab. 1 AUC value of BSSEL-LS with multiple penalty parameters in simulation

γ_B	γ_W	$u/l = 0$	$u/l = 5$	$u/l = 10$	$u/l = 15$
0	0	0.5719	0.5719	0.5719	0.5719
0	10^{-1}	0.5791	0.5846	0.5964	0.6095
10^{-3}	0	0.5801	0.6288	0.6416	0.6422
10^{-3}	10^{-1}	0.5809	0.6368	0.6473	0.6501

2.2 实证分析

在收集信用卡客户数据过程中，存在大量的新用户信息，由于其入库时间较短，数据库无法为客户

精准地标记其是否违约，因此实证分析中将此类真实的、没有标签信息的数据作为无标签数据. 随机选取 $l=4200$ 个有标签数据，“好”“坏”客户比为 20∶1，依次选取 $u=2100,4200,10500$. 对无标签数据标签化处理后，将两类数据的多数类样本分别随机分成 20 组，考察增加无标签数据后 BSSEL 方法的效果.

表 2　实证中 BSSEL-LS 在多重惩罚参数下的 AUC 值

Tab. 2　AUC values of BSSEL-LS with multiple penalty parameters in empirical test

γ_B	γ_W	$u/l=0$	$u/l=0.5$	$u/l=1$	$u/l=2.5$
0	0	0.6922	0.6922	0.6922	0.6922
10^{-3}	10^{-1}	0.7101	0.7183	0.7254	0.7428

从表 2 可以看出，随着无标签数据量的增加，多视图模型（第 3 行）表现越来越好，而有监督模型的结果（第 2 行）不变，且表现不如 BSSEL 模型.

表 3 结果显示，当增加的无标签数据越多，BSSEL 模型效果提升越大（第 6 行和第 7 行）. BSSEL 模型在只考虑有标签数据时效果相比其他有监督模型也较好，此外它能充分利用无标签数据，从而提高模型的准确率.

表 3　不同方法在不同的 u/l 比例下的 AUC 值

Tab. 3　AUC values of methods with different u/l

	$u/l=0$	$u/l=0.5$	$u/l=1$	$u/l=2.5$
LDA	0.6955	—	—	—
Naive Bayes	0.5482	—	—	—
LR	0.6483	—	—	—
SVM	0.6758	—	—	—
BSSEL-LS	**0.7101**	**0.7183**	**0.7254**	0.7428
BSSEL-SVM	0.7097	0.7172	0.7245	**0.7456**

3　结论

本文在对非均衡无标签数据构造分类器时，基于多视图半监督方法和集成学习的思想，提出了 BSSEL 方法. 该方法通过对无标签数据标签化和均衡化处理，解决了数据分组问题，再与分组的有标签数据结合，构造多组子分类器，最后的结果由子模型取均值得到. 其原理简单易懂，方法易于实现，且降低了模型预测的方差. 模拟实验和实证分析的结果显示，该方法在处理非均衡数据和无标签数据时具有良好的表现.

本文在算法实现过程中，再生核选用了高斯径向基核，若考虑不同核函数对应的数据结构不同，采用多个核函数的组合形式，模型效果可能会有提升. 此外，多视图个数采用对变量进行聚类的方式得到，可以考虑其他方式选取视图个数，且针对不同问题构造不同的结合算子 C，可能会改进模型分类效果.

参考文献（References）

[1] ZHU X. Semi-supervised learning literature survey [R]. Madison, WI: Department of Computer Sciences, University of Wisconsin-Madison, 2005.

[2] 周志华. 半监督学习中的协同训练风范[C]//机器学习及其应用. 北京：清华大学出版社，2007.

[3] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training [C]//COLT '98: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. New York: Association for Computing Machinery, 1998: 92-100.

[4] SUN X Y, GONG D W, ZHANG W. Interactive genetic algorithms with large population and semi-supervised learning [J]. Applied Soft Computing Journal, 2012, 12(9): 3004-3013.

[5] ZHOU D, BOUSQUET O, WESTON J, et al. Learning with local and global consistency[C]// NIPS '03: Proceedings of the 16th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2003: 169-176.

[6] BELKIN M, NIYOGI P, SINDHWANI V, et al. Manifold regularization: A geometric framework for learning from examples[J]. The Journal of Machine Learning Research, 2006, 7: 2399-2434.

[7] OLIVER A, ODENA A, RAFFEL C, et al. Realistic evaluation of deep semi-supervised learning algorithms [C]// NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2018: 3239-3250.

[8] WESTON J, RATLE F, MOBAHI H, et al. Deep learning via semi-supervised embedding[C]// Neural Networks: Tricks of the Trade. Berlin: Springer, 2012: 639-655.

[9] CALMA A, REITMAIER T, SICK B. Semi-supervised active learning for support vector machines: A novel approach that exploits structure information in

data[J]. Information Sciences, 2018, 456: 13-33.
[10] DRUGMAN T, PYLKKONEN J, KNESER R. Active and semi-supervised learning in ASR: Benefits on the acoustic and language models[DB/OL]. [2019-04-01]. https://arxiv.org/abs/1903.02852.
[11] MUSLEA I, MINTON S, KNOBLOCK C A. Active + semi-supervised learning = robust multi-view learning [C]// ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2002: 435-442.
[12] SUN S. A survey of multi-view machine learning[J]. Neural Computing and Applications, 2013, 23(7): 2031-2038.
[13] HOTELLING H. Relations between two sets of variates[C]// Breakthroughs in Statistics. New York: Springer, 1992: 162-190.
[14] SUN S. Multi-view Laplacian support vector machines [C]// Advanced Data Mining and Applications. Berlin: Springer, 2011: 209-222.
[15] MINH H Q, BAZZANI L, MURINO V. A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning [J]. The Journal of Machine Learning Research, 2016, 17(1): 769-840.
[16] 徐蓉,姜峰,姚鸿勋.流形学习概述[J].智能系统学报, 2006,1(1): 44-51.
[17] BROUARD C, D'ALCHÉ-BUC F, SZAFRANSKI M. Semi-supervised penalized output kernel regression for link prediction. ICML '11: Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison, WI: Omnipress, 2011: 593-600.
[18] MINH H Q, SINDHWANI V. Vector-valued manifold regularization[C]// ICML '11: Proceedings of the 28th International Conference on Machine Learning. Madison, WI: Omnipress, 2011: 57-64.
[19] MICCHELLI C A, PONTIL M. On learning vector-valued functions[J]. Neural Computation, 2005, 17(1):177-204.
[20] MROUEH Y, POGGIO T, ROSASCO L, et al. Multiclass learning with simplex coding [C]// Advances in Neural Information Processing Systems 25 (NIPS 2012). Lake Tahoe, NEV: Neural Information Processing Systems Foundation, 2012.
[21] VIAENE S, DERRIG R A, DEDENE G. Cost-sensitive learning and decision making for massachusetts pip claim fraud data[J]. International Journal of Intelligent Systems, 2004, 19(12): 1197-1215.
[22] MAJID A, ALI S, IQBAL M, et al. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines [J]. Comput Methods Programs Biomed, 2014, 113(3): 792-808.
[23] WU T T, LANGE K. Multicategory vertex discriminant analysis for high-dimensional data[J]. Annals of Applied Statistics, 2010, 4(4): 1698-1721.
[24] 李晓刚.个人信用风险评估的一种基于 XGBoost 的集成学习方法[D].合肥:中国科学技术大学,2018.
[25] BARTLETT P L, MENDELSON S. Rademacher and Gaussian complexities: Risk bounds and structural results [J]. The Journal of Machine Learning Research, 2002, 3: 463-482.

(上接第 595 页)

[22] POPOV P P, WANG H, POPE S P. Specific volume coupling and convergence properties in hybrid particle/finite volume algorithms for turbulent reactive flows [J]. Journal of Computational Physics, 2015, 294: 110-126.
[23] BARLOW R S, FRANK J H. Effects of turbulence on species mass fractions in methane/air jet flames[J]. Symposium (International) on Combustion, 1998, 27: 1087-1095.
[24] ANKARAN R S, HAWKES E R, CHEN J H, et al. Structure of a spatially developing turbulent lean methane-air Bunsen flame[J]. Proceedings of the Combustion Institute, 2007, 31: 1291-1298.
[25] BROWN P N, BYRNE G D, HINDMARSH A C. VODE: a variable coefficient ODE solver[J]. SIAM Journal on Scientific and Statistical Computing, 1989, 10: 1038-1051.