三重集约束下的自适应 SSLboost 分类方法 *

原 鑫，王振友†

(广东工业大学 应用数学学院，广州 510520)

摘 要：在分类问题中，常用的高效算法有半监督学习算法、bagging 算法和 boosting 算法等，当标记数据很少、数据间差异较大时，很难找到有效的规则来分类。针对此问题提出了三重集约束下的 boosting 分类算法，对标记数据、伪标记数据、无标记数据进行三重约束划分，同时引入平衡函数，将更新数据的近邻两点加权，确立数据空间稳定点，根据稳定点信息对分类器进行迭代，采用梯度下降法使得平衡函数收敛，得到最终的伪标记数据和分类器。经过 UCI9 个数据集的实验，验证了该算法更为高效、可行。

关键词：boosting 算法；空间稳定点；三重集约束；平衡函数

中图分类号：TP301.6 doi: 10.19734/j.issn.1001-3695.2019.03.0037

Adaptive SSLboost classification algorithm based on triple set constraint

Yuan Xin, Wang Zhenyou†

(*Dept. of Applied Mathematicslege, Guangdong University of Technology, Guangzhou* 510520, *China*)

Abstract: In the classification problem, the commonly algorithms are Semi- supervised learning algorithm bagging and boosting classification algorithm. When there is little mark data and the difference between data is large, It's hard to find effective regulations to classify them. A boosting classification algorithm under triple set constraints aims at solving this problem, The algorithm divides the mark data, pseudo mark data and unmarked data by triple constraints, and introduces a balance function to weight the two neighbors of the mark data and find the stable data for this space, then iterates the classifier according to that data information, Through the gradient descent method, the balance function converges and finally get the pseudo-marker data and classifier. Experiments on nine data sets of uci verify that this algorithm is more efficient and feasible.

Key words: boosting algorithm; space stability point; triple set constraints; balanced function

0 引言

现实生活中往往更容易收集到大量的无标记样本，获取标记的过程中需耗费大量的人力物力，这种现象在互联网应用中更为明显，在进行网页推荐时，通常希望通过用户手动标记来获取最可靠的用户信息，然而很少用户会消耗时间来完成信息的填写,大量的无标记样本给分类造成一定的困难，研究人员希望通过少量的标记样本来获取到大量无标记样本的重要信息。对此需要利用无标记样本特征，使得无标记样本所揭示的数据分布信息与类别标记相联系，常见的方法有半监督分类方法[1]、bagging 方法和 boosting [2]方法。

由 Freund 和 Schapire 提出的 Adaboost 算法，是最初的 boosting 分类方法，当前最先进的一种 boosting 算法是 MSSBoost 算法[3]，采取了成对约束优化来提高基分类器的分类性能，通过在标签数据上训练模型，使用该模型来预测无标记数据的标签，创建伪标签，将伪标记指派给无标签数据，在每一轮的迭代预测中，尽可能的减少经验损失，以此迭代预测无标记数据类标签。另一种将 boosting 与半监督学习相结合的算法是 Semiboost 算法[4]，旨在训练过程中充分使用流行和聚类假设[12]，它是一种典型的无约束条件下的半监督 boosting 方法。对比以上成对约束与无约束的方法，提出了一种三重集约束下的 SSLboost 分类算法，将流形空间[5]算法的内部线性表示扩展到流形外部约束，确立空间稳定点，在平衡函数约束下，迭代收敛生成数据空间伪样本的方式，来训练得到新的分类器，每一轮迭代中充分考虑到更新后的数据空间结构变化，从而有效的解决有标记样本少，识别度低的情况。对于不平衡的类分布，Boosting 算法的难点在于没有保证有一个最优的联合分类器或概率估计，因此很可能增加分类的复杂性，对此考虑从数据分布层面来解决。

简言之 SSLboost 算法的优点有:a)根据无标记数据集较大时效率不足的缺点，通过对伪标记数据权值自适应调整，得到能够更为精确预测无标签数据的分类器与流形空间模型；b)经过 UCI 数据集上的分类实验，针对不平衡数据的分类情况，在不同的基分类器下分类效果取得了一定的提升，与之前的 Mssboost 、Semiboost 分类方法相比，由原来的每次迭代计算最小化相似度损失函数，转变为迭代寻找数据空间稳定点，得到刻画数据真实空间的伪标记数据，减少了更新数据相似度的环节，为 boosting 算法提供了一种获取数据真实空间结构的方法，在每次迭代预测数据类标签的过程中，得到独立于分类器的数据；c)在三重集条件的约束下，采取不同的数据权值更新策略，建立平衡函数，为更新过程提供了可行的终止条件。

1 相关工作及算法思想

1.1 数据处理的相关工作

机器学习中，解决不平衡数据问题常采用三种策略：代

收稿日期：2019-03-04；**修回日期：**2019-05-05 **基金项目：**广州市科技计划项目（201707010435）

作者简介：原鑫(1994-)，男，山西长治人，硕士研究生，主要研究方向为算法设计与数据挖掘、机器学习；王振友(1979-)，教授，主要研究方向为机器学习、优化计算、医学统计分析(zywang@gdut.edu.cn).

价敏感策略、采样策略和主动学习策略[7]。常见的有代价敏感 SVM 、 KELM 算法[8] 、 Smote 采样[17]等。

代价敏感策略关注于样本的重要性，根据重要性确定样本的误分代价来分类。其中代价敏感 SVM [6]通过对多数类样本和少数类样本赋予不同的代价损失，引入一个权值偏离的支持向量机，使得最终的超平面偏离少数类样本，这种多次计算误分代价的方式，容易受到噪声点、离群点的影响，从而产生过拟合现象。

KELM 算法通过对多数类的样本随机抽样实现均衡训练，根据不均衡程度确定训练次数实现分类。目前最先进的有 UBKELM [15]，该方法不需要对样本分配权重，而是将不均衡问题[18]划分为几个平衡问题进行讨论，使用高斯核 ELM 分析平衡分类问题的分类模型，由于基分类器在平衡数据上实现分类，因此 KELM 算法取得了不错的效果。

主动学习相对以上两种方法的优势在于保证了模型整体在相对平衡的数据空间上进行，不需要对数据做过多的划分，但由于以往的规则过于局限，获取数据真实空间与停止标准设定是一大难题。由- Huang 等人提出了一种利用聚类信息和分类间隔的样例选择主动学习策略[16]，解决了采集孤立点的问题，使得采集到的样本更具代表性，但在数据真实空间的确定上效果不足，为了克服以上主动学习所存在的局限性，本文通过引入伪标记点来平衡标签数据与未标签的数据空间，确定数据空间强相关点与弱相关点，使无标记数据通过伪标签信息携带更多与有标记数据集相关的信息，并考虑部分邻域中的局部约束来逼近这些完全约束。

Smote 算法的核心思想是对于少数类样本进行分析，从少数类样本的 K 个近邻中选择 N 个样本，采取随机线性插值的方法，与原先的标记数据组成新的训练集，进一步对无标记数据标签进行预测，常见的有 BSMOTE [13]、 MWMOTE [14]等，由于 Smote 算法在采样过程中存在一定的盲目性，无法从根本上解决不平衡数据的分布问题，若少数类别样本处在样本分布的边界位置，由此衍生出的样本会造成数据的边缘化，使得类别间隔越来越模糊。针对此问题本文提出确定数据空间的稳定点，在相关性研究的基础上，得到反映数据空间结构的中心点与边界点，减小数据边缘化对分类的影响。

1.2　三重约束分类总体思想

三重约束思想[9]之前应用于度量矩阵的学习，在该思想的启发下，在算法层面上提出三重约束下解决分类问题的方法，首先将样本分为标记样本和无标记样本，记标记样本 $X_L=\{X_1,X_2,\ldots X_l\}$，无标记样本 $X_U=\{X_{l+1},X_{l+2},\ldots X_u\}$。在有标记样本中找出与所取样本距离相近的前 K 个样本，取该样本的加权距离生成伪标记样本，并赋予相同的标记，为了方便三重约束集的建立，选取数据的两个近邻点，其余的权值设为零，即 $X_k=w_1*X_{k\_nearest1}+w_2*X_{k\_nearest2}$，这时需要考虑到新增样本加权求和之后样本与之前标记样本重合的情况，如果该样本增加之后的标记与原先样本标记存在差异，会给分类带来一定的偏差，由于重叠数据包含不同类的概率几乎相同的区域，因此传统的分类器很难找到一种可行的分类方法。为了更好地解决这个问题，可以通过给分类样本设立平衡函数，来均衡伪标记数据在总样本中的分布，来达到良好的分类效果。

在众多不同的度量学习中，本文选择 Mahalanobis 距离度量[10]。 Mahalanobis 距离度量可以看做是对输入空间线性变换后的欧氏度量。形式上，两个例子 X_i 和 X_j 之间的 Mahalanobis 距离被定义为

$$\rho_M^2(x_i,x_j)=(x_i-x_j)^T M(x_i-x_j) \tag{1}$$

本质目标是通过确定输入特性的重要性及其相关性来了解示例之间的不同之处。对此取标签数据中每个数据点相邻排序的前两个点，取两个邻近点的加权和作为与原标记相同的点，通过平衡两个数据点的权值，使得第二个相近点携带更多与无标记样本有关的信息，同时使其减少数据空间结构改变导致的信息丢失。

对上述三种样本建立三重约束，在最大间隔原理的启发下，对于三重约束，通过寻找合适的马氏度量使得不等式满足

$$\rho_M^2(x_j,x_k)\le\rho_M^2(x_i,x_k) \tag{2}$$

下面的权值损失函数惩罚了对上述不等式的小违反，定义为

$$w_2=\frac{1}{\mathrm{m}}\sum_{i=1}^{m}\sum_{j=1}^{m}[1+\frac{\rho_M^2(x_i,x_k)-\rho_M^2(x_j,x_k)}{m}]*$$

$$[z]*:=\begin{cases}z & z>0\\ \max(z,0) & z\le 0\end{cases}$$

$$w_1=\frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{m}W_{ik}\rho_M^2(x_i,x_k)=\frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{m}W_{ik}\|x_i-x_k\|^2M$$

$$=\frac{1}{2}tr(XL^TX^TM) \tag{3}$$

m 是随机抽样所选取的样本数，对于任意的 $\mathrm{i,j,k}\in L,U,k$，设置惩罚项的目的在于：使得与无标记数据相关的权值在限度内尽可能地大。w_1 使得伪标记数据与标记数据尽可能距离相近，w_2 则迫使训练的伪标记数据与无标记数据具有相似的标签信息。由于以下相关度度量介于 0 到 1，可以定义：

$$W_{ik}=\begin{cases}1, & \text{if } \mathrm{e}^{-(x_i-x_k)^2}>0.5\\ 0, & \text{else}\end{cases}$$

$$W_{jk}=\begin{cases}1, & \text{if } \mathrm{e}^{-(x_j-x_k)^2}>0.5\\ 0, & \text{else}\end{cases} \tag{4}$$

为了求得适应该数据流形空间的马氏矩阵，在这里采取梯度下降的方法来寻找合适的度量矩阵，并且定义：

$$M(t)=(L(t)-D(t))*(L(t)-D(t))^T$$

时间 t 表明每个时刻下的度量矩阵变化，W_{Lk} 与 W_{Uk} 决定每个时刻下的邻接矩阵 $L(t)$ 的变化， $\mathrm{D(t)}$ 由邻接矩阵每一列值加于主对角线上得到，并确保了马氏矩阵的半正定性，进而转换为半正定规划问题。

1.3　平衡函数及参数选取

在图半监督学习的启发下，引入邻接矩阵来分析每次迭代加入伪标记数据后的数据空间相似度变化，具体方法及参数选择如下：

a)每次迭代选取与标记样本相同大小的伪标记数据，计算出每个样本之间的相似度，根据式(4)得到 W_{ik}，方便计算 w_1、w_2。

b)在无标记样本中抽出与伪标记样本相同大小的数据集。取样次数记作 N，根据式(4)得到最终的邻接矩阵。

c)求出无标记样本与伪标记样本之间 N 次平均后的马氏矩阵。

假设无标记样本数量为 n，在这里 N 为平衡函数收敛时候的终止次数，目标是对近邻点之间建立平衡函数，求得有标记数据和无标记数据的权值，找到最佳反映数据流行空间结构的伪标记数据，而这样的平衡函数是基于无标记样本、有标记样本、伪标记样本之间三重约束的条件来自适应分析的。基于以上思想定义平衡函数,其中平衡参数 α 在 0到1 之间步长为 0.1 取值，平衡函数如下：

$$G(t)=\min\ tr(XL^TX^TM)+\frac{\alpha}{\mathrm{m}}\sum_{i=1}^{m}[1+\frac{\rho_M^2(x_i,x_k)-\rho_M^2(x_j,x_k)}{m}]*$$

根据文献[11]证得该平衡函数是收敛的，并且收敛速度较快，w_2 中参数值保证了权值大小在一定的区间内，其中 ρ 值的大小由每个时刻的 M 值所决定。

1.4　自适应数据结构分析

在 1.3 节中，每次选取和标记样本数量相同大小的无标记数据，以此为先行条件，进一步地找到这些数据中的强相关点和弱相关点，从而保证每次选择数据之后的空间结构稳定性。

1.4.1 无标记样本的选取方法

定义相似性为

$$S(i,j)=e^{-(x_i-x_j)^2} \tag{5}$$

i,j 为选取的样本点 ID，公式的含义为样本间距离越大，相似度越低。

制定策略：假定无标记数据量为 n，对伪标记数据编号{1,2,3,…m}，在 1.2 中，每次抽出大小为 l 的无标记数据集量，同时拒绝其余样本，计算标签为 i=1,2,3…m 的样本与所取 l 个样本的相似度，找到相似度最高和最低的样本，并在剩余样本中选取下一个比目前为止看到相似度更高和更低的两个样本作为最佳样本，否则将第一次找到的样本作为最佳样本，该策略的目的在于找到最佳的 l 值，来保证每次选择的无标记数据存在一定量的强相关数据和弱相关数据，可以证明存在一个阈值，并且在该阈值条件下，所得到的最佳样本概率不会下降。

假设 H_best 和 L_best 为两个最佳样本点，在预定选取的无标记样本点中，最佳样本的顺序 ID 为 k，记录 R(n,l) 为成功选择最佳样本点的事件，采取的策略即排除第一个 l-1 人，找到最终的最佳样本，可以得到

$$P(R_{n,l}|H\_best=k)=\begin{cases}\dfrac{1}{l-1}, & k\in\{1,2,3...l\text{-}1\}\\ \dfrac{l-1}{k-1}, & k\in\{l,l\text{+}1,...n\}\end{cases}$$

对于公式的解释为：如果最优样本为前 $l-1$ 个样本之一，那么样本选取概率为恒定值，当最优样本落在剩余无标记样本中时，由于最佳样点 ID 为 k，所以当且仅当 l-1 中的一个在最初的 k-1 个样本点中是最佳样本时，该策略成立。

$$P(R_{n,l})=\sum_{k=1}^{n}P(R_{n,l}|H\_best=k)\times P(H\_best=k)$$

$$=\sum_{k=1}^{l-1}\frac{1}{l-1}\times P(H\_best=k)+\sum_{k=l}^{n}\frac{l-1}{k-1}\times P(H\_best=k)$$

$$=\frac{1}{n}+\sum_{k=l}^{n}\frac{l-1}{k-1}\times\frac{1}{n}=\frac{1}{n}+\frac{l-1}{n}\sum_{k=l}^{n}\frac{1}{k-1}$$

由于现实生活中存在着大量的无标记样本，且无标记样本点数量远多于有标记样本点，因此当 n 足够大时，选择最优样本的概率转变为以下形式：

$$P(R_{n,l})\to t\int_t^1\frac{1}{x}\,dx=-t\times\ln t\quad n\to\infty$$

上式成立条件为：$t=\dfrac{l}{n}$

对上式求导 $\dfrac{d}{dt}(-t\ln t)=-\ln t-1$

即 $t=e^{-1}$ 时 $P(R_{n,l})$ 最优。

$$l=e^{-1}\times n\approx 0.37n$$

因此每次抽取 [0.37n] 个样本点，在该策略下进一步找到强相关数据点和弱相关数据点，即 H_best 和 L_best 两个最佳样本点。

1.4.2 空间稳定点确定方法：

将每次得到的强相关数据点和弱相关数据点加入到所抽取的样本点中，由式(5)得到

$$S(H\_best,X\_mean)=e^{-(H\_best-X\_mean)^2}$$

找到满足

$$S(i,j)<=S(H\_best,X\_mean) \tag{6}$$

$$S(i,j)<=S(X\_mean,L\_best) \tag{7}$$

的数据点，将其命名为数据空间结构的稳定点，与所抽取样本数据组成 m 个点共同作用于马氏矩阵的

迭代计算，优点在于每次所选取的数据很好地刻画了数据的中心、边缘性质，数据流程图如图 1 所示。

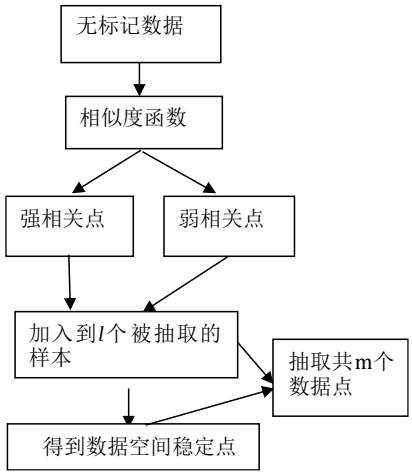

图 1　无标记样本抽取流程图

Fig. 1　Unlabeled sample extraction flowchart

1.4.3 空间稳定点示例说明：

不失一般性，抽取两组二维正态分布数据：

$$a=\{N_1(\mu_1,\sigma_1)\,|\,\mu_1\in[1,5],\sigma_1\in[1,2]\}$$

$$b=\{N_2(\mu_2,\sigma_2)\,|\,\mu_2\in[6,10],\sigma_2\in[1.5,3]\}$$

μ_1 分别代表类别一数据中两组特征的均值，σ_1 分别代表类别一数据的方差，μ_2 代表类别二两组特征的均值，σ_2 代表类别二数据两组特征的方差。

数据产生于不同的数据区间，实验验证由强相关数据和弱相关数据引导的数据采样大多集中在中心点与边界点，且每次采样后的数据信息关联于采样数据外的数据信息，以此计算得到有效的马氏矩阵来分析伪标记数据内部的加权结果。

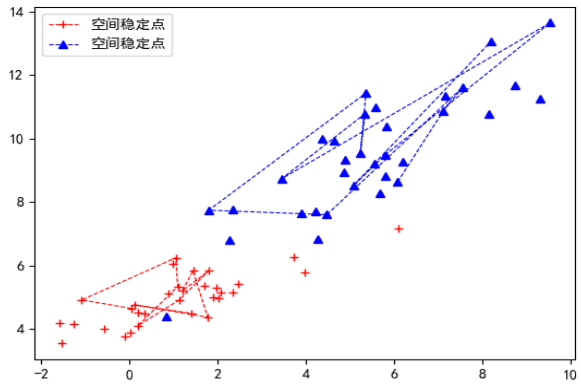

图 2　数据空间稳定点的确定

Fig. 2　Determination of data space stability points

图 2 为 a 和 b 两组数据下的样本分布，在图中利用式(6)(7)对两类样本点中的稳定点做标记，两组实验数据表明，由强相关数据和弱相关数据引导的数据采样大多集中在数据的中心点与边界点，最终建立伪标记数据与无标记数据的相关性

质，使得每次采样后的数据更具代表性。

空间稳定点的选取建立在找到合适的强相关数据点与弱相关数据点的基础之上，在上述的策略下，对应于 H_best 和 L_best 两个最佳样本点的选取。也就是说，强相关数据与弱相关数据的代表性决定了稳定点对于数据空间的贡献大小。

式(6)(7)使得基于强相关点和弱相关点的取值落在一定的区间内，稳定点能够恰当取在数据空间的中心点与边界点处，从而有效地反映数据的真实空间。

在得到数据空间稳定点的过程中，将策略下的两个最佳样本点加入到所抽取的 l 个样本中，因此每次抽取的样本不应当过小，否则需要多次采样来满足实验的需求，在该策略下的样本抽样大小很好的适应了实验迭代次数较少的情况。

对于实验策略下的最佳样本点选取过程，最佳样本点一定概率上采样于抽取的数据中，也可能取自抽样的样本点之外，在这种策略下很好的保证了样本选取的多样性，通过局部信息逼近于整体信息。

2　算法流程及伪代码

2.1　算法流程图

本文的算法流程图如图 3 所示。

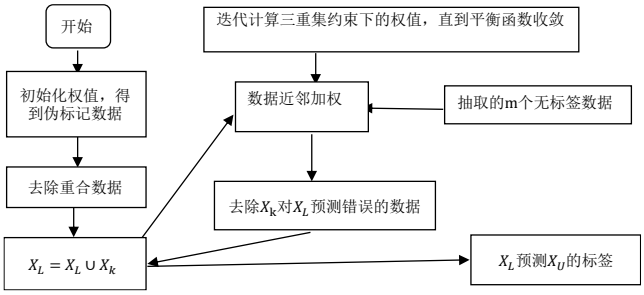

图 3　算法流程图

Fig. 3　Algorithm flowchart

2.2　算法伪代码

1. $Input\ data\{X_L, X_U\}$、$step$　输入标记数据与未标记数据 $X = X_L \cup X_U$

2.PCA 初始化 $M(t)$，标记样本 X_L 大小为 m

3 then 在 X_U 中找出 X_L 的每个近邻点序号　##

4.for k in range(m) do:

$$X_k = w_1 * X_{k\_nearest1} + w_2 * X_{k\_nearest2}$$

数据近邻两个点加权和

5. X_k ← 去除与标记点重合的伪标记点(数据很大时必要的步骤)

6.在 X_k 和 X_L 上训练分类器 C

X_k ← 去除分类器 C 下伪标记数据预测与原有标记存在差异的点

7. 求得平衡函数的梯度 $\nabla G(t)$　　$X_k = X_L \cup X_k$

$M(t+1) = M(t) - step * \nabla G(t)$

8. $X_L = X_L \cup X_k$　　更新 w_1、w_2

X_U = random.sample (X_U ,m)

9.if G(t) converge　使用 X_L 预测 X_U 中每个数据的标签，得到该数据分布下的平均准确率

10.else 执行##

11.break 语句跳出循环

3　实验过程与结果分析

3.1　实验概述

实验在 9 个 UCI 数据上进行，并报告了每五十次求出结果的平均准确率，将实验结果与文献[4]的 Mssboost 算法作比较。实验对有标记样本采取随机抽样策略，抽取 10% 的数据作为带标记样本，剩余数据去掉标签，使用上述算法思想，抽取相同大小

的无标记数据和迭代更新后的标记样本，加权得到伪标记数据，循环更新分类器，对此用三种不同的基分类器来验证三重约束条件下的 SSLboost 算法的适应性，分别取 SVM、Naive-Bayes、ANN 作为基分类器对实验结果进行分析，以 SVM 为基分类器时，对比分析 SSLboost 算法和 Mssboost 分类效果的同时，加入了 LapSVM 算法[5]进行比照。并且进一步对比分析了在 Naive-Bayes 分类器下的 SSLboost 与 AdaBoost [20]的实验效果。

3.2　实验数据选取与分析：

选取 UCI 上的 9 个数据作出分类对比，数据的信息如表 1 所示。

表 1　数据信息表

Table 1　Data information table

数据名称	数据大小	特征数目	类别数目
Iris	150	4	3
Balance	625	4	3
Glass	214	9	6
Vowel	990	14	11
Sonar	208	60	2
banknote	1372	4	2
wine	178	13	3
zoo	101	18	7
Magic	19020	11	2

不同基分类器下的实验效果如图 3 所示。

图 3 是 Balance 数据和 Zoo 数据在两种基分类器 SVM、Naive-Bayes 下的数据分类对比效果， Balance 与 Zoo 数据为典型的多类别不均衡分布数据集，在文献[7]中，给出了该数据集的不平衡度，由此来验证 SSLboost 算法在不均衡分布数据集上的分类效果。

三重集约束下的 boosting 算法的优势在于获取了真实的数据空间结构，对于类别数目较少的数据，结合数据之间的关联性，迭代生成了更多与该类别相关的伪标记数据，为有标记数据和无标记数据的加权提供了有效的阈值边界，由此增加了对于类别较少的数据的判断依据，因此在较少类别数据的识别上可以取得良好的效果。

3.3　实验结果及分析

分别在三种不同基分类器下进行分类实验，对每组数据进行准确率对比，在不平衡度较大的数据集上取得了良好的实验效果，实验中基分类器 SVM 的惩罚系数为 1.0，核函数采用 RBF 核，单层 ANN 激活函数采用 relu，优化器设置为 Adam。在标记数据占比 10% 的参数环境下验证算法的适应性与稳定性，并对不同均衡度的数据集在 5%、10%、20%、30%、40% 有标签数据集占比情况下对比分析了标记数据大小对于算法的影响，得到的实验结果如表 2~4 所示。

综合三种基分类器下的实验结果，三重集约束下的 SSLboost 算法较之前的算法效果均得到了一定程度的提升综合三种基分类器下的实验结果，三重集约束下的 SSLboost 算法较之前的算法效果均得到了一定程度的提升。经过 9 个数据集上的对比实验，可以清晰看到在 Glass 数据集上准确率提升最为显著，其中 Glass 数据集的不均衡度为 6.38，Balance 数据不均衡度为 11.78，ANN 作为基分类器时对 Balance 数据集的提升较高。

另外，实验选取两种不平衡度较高的数据集，对于不同有标记数据占比情况，分别在 SVM 基分类器、Bayes 基分类器、ANN 基分类器下进行了实验对比。实验效果如图 4 所示。

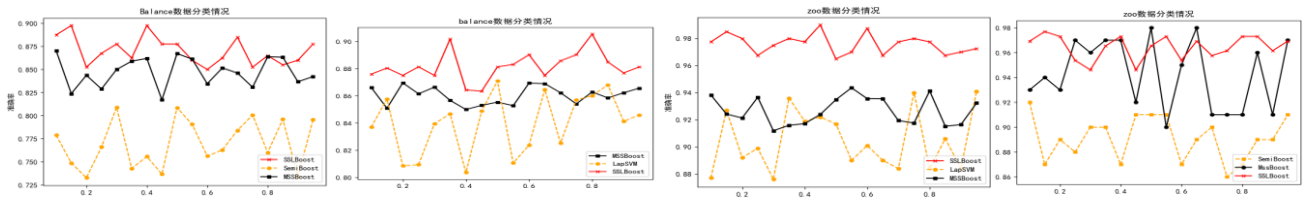

图 3　不同基分类器下实验效果对比

Fig. 3　Comparison of experiment results under different base classifiers

表 2　SVM 基分类器下的 10%标记数据实验

Table 2　10% label data Experiment of SVM classifier

数据	算法			
	LapSVM	Semiboost	MSSboost	SSLboost
Iris	93.45±4.5	86.19±9.7	97.01±1.9	97.82±1.2
Balance	84.32±4.1	83.84±3.4	86.01±1.1	89.90±3.5
banknote	97.22±1.3	97.4±2.1	98.13±1.5	98.6±1.1
Wine	93.40±3.2	96.49±3.0	95.60±3.4	95.97±3.2
Zoo	91.40±3.8	92.0±3.3	92.79±1.6	93.12±1.1
Magic	76.20±3.6	75.2±2.7	76.5±3.1	80.62±1.3
Glass	56.8±4.1	52.94±4.6	60.70±3.9	97.23±1.7
Vowel	40.34±2.1	40.50±2.2	39.40±3.8	50.17±2.5
Sonar	69.10±2.8	70±4.6	71.50±2.3	75.91±1.9

表 3　Bayes 基分类器下的 10%标记数据实验

Table 3　10% label data Experiment of Bayes classifier

数据	算法			
	Adaboost	Semiboost	MSSboost	SSLboost
Iris	89.06±5.3	90.47±6.3	94.9±4.2	96.01±2.1
Balance	62.51±3.6	77.36±4.1	84.03±3.1	85.71±2.0
banknote	92.3±3.1	94.5±2.3	96.5±1.7	97.46±1.3
Wine	89.66±6.4	92.28±2.6	96.9±3.1	97.7±1.9
Zoo	86.39±3.5	89.52±3.8	94.2±3.8	96.1±1.6
Magic	69.5±8.0	76.3±4.7	76.85±3.6	85.36±1.8
Glass	49.35±4.8	56.51±7.1	59.90±2.1	87.3±1.9
Vowel	40.23±0.9	44.5±2.2	51.90±2.9	60.7±1.5
Sonar	66.8±4.7	70.91±5.4	70.35±5.9	71.24±4.3

表 4　ANN 基分类器下的 10%标记数据实验

Table 4　10% label data Experiment of ANN base classifiers

数据	算法			
	ANN	Semiboost	MSSboost	SSLboost
Iris	66.1±5.9	80.27±2.1	87.6±3.2	86.8±2.5
Balance	51.7±3.1	70.51±5.9	72.36±4.2	80.23±3.2
banknote	96.13±0.9	97.01±0.7	97.28±0.2	98.9±0.3
Wine	92.6±4.2	90.45±3.6	93.14±2.7	94.3±3.9
Zoo	67.59±3.7	88.7±3.1	91.6±4.3	96.7±2.7
Magic	60.5±9.1	70.69±2.3	71.9±1.1	79.28±4.3
Glass	44.32±1.7	61.32±4.7	62.1±1.9	97.8±1.3
Vowel	32.6±4.5	41.26±2.4	43.5±2.9	57.1±1.4
Sonar	59.40±6.3	70.12±5.3	72.33±3.6	76.52±1.2

经过以上实验得到当前常见算法及最先进算法在 UCI 数据上的平均准确率，从而验证了 SSLBoost 算法在不同占比有标签数据上的算法稳定性，通过不同参数环境下的实验效果对比，验证了该算法的提升效果。

4　结束语

三重约束下的 SSLboost 算法将数据三重约束划分，在半正定规划问题中平衡函数的约束下，确定迭代过程的空间稳定点,确立主动学习模型，相比之前的 boosting 算法在数据结构的获取方面有所突破，有效联系了图论中拉普拉斯矩阵来分析实现每一轮迭代计算，并在高斯分布和不均衡分布下的数据集上均取得很好的效果，在今后的实验中将采取不同的相似度函数来对比实验，尽量避免重复实验来确定平衡参数 α 与 step 的取值。

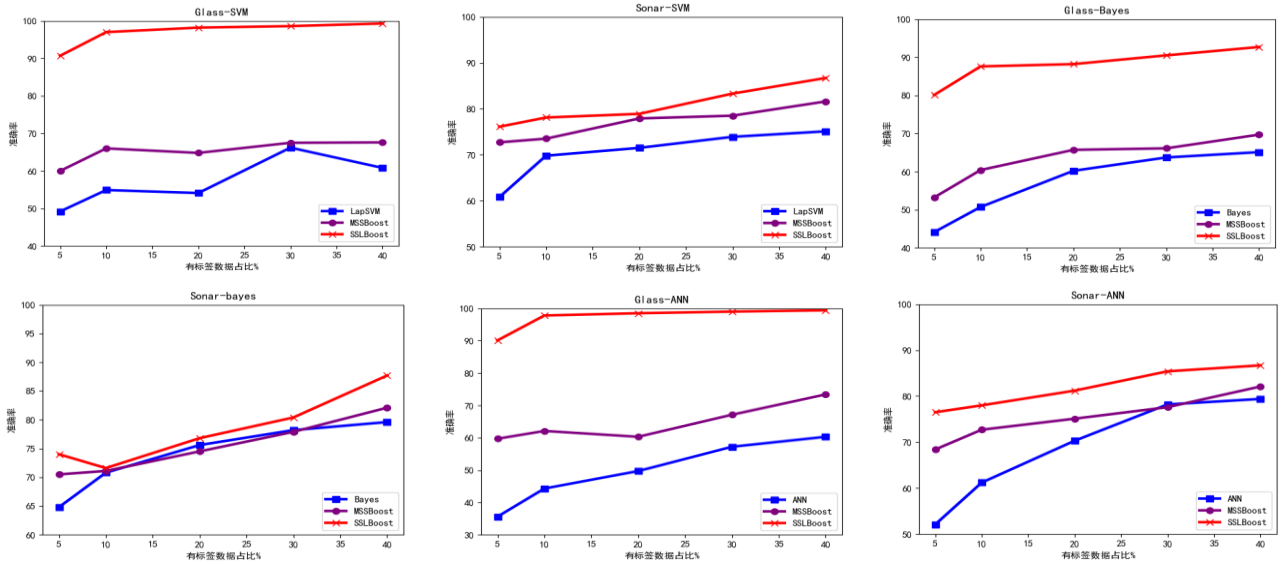

图 4　不同占比标签数据下三种基分类器的平均准确率

Fig. 4　Average accuracy of three base classifiers under different proportions of labeled data

参考文献：

[1] Yi Wang, Tao Li. Improving semi-supervised co-forest algorithm in evolving data streams [J]. Applied Intelligence, 2018 (48): 3248-3262.

[2] Freund. Y, Schapire. R. E. A decision-theoretic generalilization of on-line learning and an application to boosting [C]//Proc of European

Conference on Computational Learning Theory. 1995: 119-139.

[3] Jafar T. MSSBoost: a new multiclass Boosting to Semi-Supervised learning [J]. Neurocomputing, 2018 (6): 251-266.

[4] Mallapragada Pawan Kuma, Jin Rong. Jain Anil K, *et al.* SemiBoost: Boosting for semi-supervised learning [J]. IEEE Trans on Pattern Analysis &Machine Intelligence, 2009, 31 (11): 2-000-2014.

[5] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006 (7): 2399-2434.

[6] Cheng Fanyong, Zhang Jing, WenCuihong. Cost-Sensitive Large margin Distribution machine for classification of imbalanced data [J]. Pattern recognition Letters, 2016, 80 (C): 107-112.

[7] Oh S, Lee Minsu. Zhang B T. Ensemble learning with active example selection for imbalanced biomedical data classification [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2011 (8): 316-325.

[8] Raghuwanshi B S, Shukla S. Class imbalanced learning using under bagging based kernelized extreme learning machine [J]. Neuro Computing, 2018, 56 (10): 172-187.

[9] Nguyena B, Morel C, De Baets B. Distance metric learning for ordinal classification based on triplet constraints [J]. Knowledge-based Systems, 2018, 142 (2): 17-28.

[10] Ying Yiming, Peng Li, Distance metric learning with eigenvalue optimization [J]. Journal of Machine Learning Research, 2012, 13 (1): 1-26.

[11] Vandenberghe L, Boyd S, Semidefinite Programming [J]. SIAM Review, 1996 (1): 49-95.

[12] Kumar N, Kummamuru K, *Paranjpe D.* Semi-suprevised-clustering with metric learning using relative comparisons [C]//Proc of IEEE International Conference on Data Mining. 2005.

[13] Han Hui, Wang Wenyuan., Mao Binghuan.. Borderline-SMOTE: a new over-sampling method in imbalanced datasets learning [C]// Proc of International Conference on Advances in Intelligent-Computing. 2005:878-887.

[14] Barua S, Md Lslam M, Yao Xin. MWMTE: majority imbalanced data set learning [J]. IEEE Trans on knowledge and Data Engineering, 2014, 26 (2): 405-425.

[15] Zhang Hongyan, Wang Haiyan, Dai Zhijun, *et al.* Improving accuracy for cancer classification with a new algorithm for genes selection [J]. BMC Bioinformatics, 2012 (11) :298.

[16] Huang Shengjun, Jin Rong, Zhou Zhihua, Active learning by querying informative and representative examples [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014 (10). .

[17] 邵良杉，周玉．一种改进过采样算法在类别不平衡信用评分中的应用 [J]. 计算机应用研究，2019, 36(6): 1683-1687. (Shao Liangshan, Zhou Yu, Application of improved oversampling algorithm in class-imbalance credit scoring [J]. Application Research of Computers, 2019, 36(6): 1683-1687.)

[18] 邹权，郭茂祖，刘扬，等．类别不平衡的分类方法及在生物信息学中的应用 [J]. 计算机研究与发展, 2010, 47 (8): 1407-1414. (Zou Quan, Guo Maozu, Liu Yang, *et al.* Classification method for class-imbalanced data and its application on bioinformatics [J]. Journal of Computer Research and Development, 2010, 47 (8): 1404-1414.)

[19] 才子昕，王馨月，徐剑，等．样本自适应的不平衡分类器 [J]. 计算机科学, 2019, 46 (1): 94-99. (Cai Zinxin, Wang Xinyue, Xu Jian, *et al.* Sample adaptive classifier for imbalanced data [J]. Computer Science, 2019, 46 (1): 94-99.)

[20] YOAV F, SCHAPIRE R E. A decision-theoretic generalization of online learning and an application toboosting [C]// Proc of European Conference on Computational Learning Theory. Berlin: Springer, 1995: 23-37.