

典型半监督分类算法的研究分析

孟 岩,汪云云

(南京邮电大学 计算机学院/软件学院,江苏 南京 210000)

摘 要:近年来,大量半监督分类算法被提出。然而在真实的学习任务中,研究者很难决定究竟选择哪一种半监督分类算法,而在这方面并没有任何指导。半监督分类算法可通过数据分布假设进行分类。为此,在对比分析采用不同假设的半监督分类典型算法的基础上,以最小二乘方法(Least Squares,LS)为基准,研究比较了基于聚类假设的转导支持向量机(Transductive Support Vector Machine,TSVM)和基于流行假设的正则化最小二乘法(Laplacian Regularized Least Squares Classification,LapRLSC),并同时利用两种假设的 SemiBoost 以及无任何假设的蕴含限制最小二乘法(Implicitly Constrained Least Squares,ICLS)的分类效果。得出的结论为,在已知数据样本分布的情况下,利用相应假设的方法可保证较高的分类正确率;在对数据分布没有任何先验知识且样本数量有限的情况下,TSVM 能够达到较高的分类精度;在较难获得样本标记而又强调分类安全性时,宜选择 ICLS,而 LapRLSC 也是较好的选项之一。

关键词:半监督分类;数据分布;聚类假设;流行假设

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2017)10-0043-06

doi:10.3969/j.issn.1673-629X.2017.10.010

Research and Analysis of Typical Semi-supervised Classification Algorithm

MENG Yan,WANG Yun-yun

(School of Computer and Software,Nanjing University of Posts & Telecommunications,Nanjing 210000,China)

Abstract:Large amounts of semi-supervised classification algorithms have been proposed recently,however,it is really hard to decide which one to use in real learning tasks,and further there is no related guidance in literature. Therefore,empirical comparisons of several typical algorithms have been performed to provide some useful suggestions. In fact,semi-supervised classification algorithms can be categorized by the data distribution assumption. Therefore,typical algorithms with different assumption adoptions have been contrasted. Specifically,they are Transductive Support Vector Machine (TSVM) using the cluster assumption,Laplacian Regularized Least Squares Classification (LapRLSC) using the manifold assumption,and SemiBoost using both assumptions,and Implicitly Constrained Least Squares (ICLS) without any assumption,with the supervised least Squares Classification (LS) as the base line. Eventually it is concluded that when data distribution is given,the semi-supervised classification algorithm that adopts corresponding assumption can lead to the best performance;without any prior knowledge about data distribution,TSVM can be a good choice when the given labeled samples are extremely limited;when the labeled samples are not so scarce,and meanwhile if learning safety is emphasized,ICLS is proposed,and LapRLSC is another good choice.

Key words:semi-supervised classification;data distribution;cluster assumption;manifold assumption

1 概 述

传统的机器学习技术分为两类:监督学习和无监督学习。监督学习只利用标记的样本集进行学习,而无监督学习只利用未标记的样本集进行学习,但在很

多实际问题中,有标记样本通常很难收集,而无标记样本很容易得到。例如,在垃圾邮件检测中,可以自动收集大量的邮件,却只有少量是标记的垃圾邮件;在生物学中,大量的未标记数据很容易得到,而对某种蛋白质

收稿日期:2016-10-13

修回日期:2017-01-19

网络出版时间:2017-07-11

基金项目:国家自然科学基金资助项目(61300165);高等学校博士学科点专项科研基金新教师类(20133223120009);南京邮电大学引进人才基金(NY213033)

作者简介:孟 岩(1992-),男,硕士研究生,研究方向为模式识别与机器学习;汪云云,博士,副教授,研究方向为模式识别、机器学习、神经计算等。

网络出版地址:cnki.net/kcms/detail/61.1450.TP.20170711.1455.054.html

的结构分析或者功能鉴定,可能会花上生物学家很多年的时间。因此,同时利用标记样本和未标记样本的半监督学习技术在近些年发展迅速^[1-4]。

半监督分类算法利用大量的无标记样本与有标记样本一同训练,以增强分类效果。为了更加有效地利用有标记样本,提出了一些数据分布假设,常见的有两种:一种是聚类假设,分类边界穿过数据低密度区域,把数据分为几簇聚类,在一簇中的样本具有相同的标签;另一种是流行假设,充分利用数据在低维空间上的流行分布,并通过拉普拉斯图构造数据流行内在的几何结构,从而在这个图中相似的样本具有相同的标签。

几乎所有的半监督分类算法都显式或隐式地利用了这两种假设^[1,4]。例如,转导支持向量机(TSVM)^[5]和其他扩展方法^[6-8]都利用了聚类假设。而那些基于图的半监督分类方法(graph cuts^[9], label propagation^[10-11])和流行正则化最小二乘法(LapRLSC)^[12]都利用了流行假设。除此之外,有的方法同时利用这两种假设来增强分类效果。半监督 Boosting^[13-14]就是一种同时利用两种假设,并利用迭代的 boosting 算法^[15]来增强分类效果的半监督分类方法。另一种相关的算法是正则化 Boosting 算法^[16],它同时利用 boosting 框架和结合了平滑性的以上两种常用假设。上述方法都显式地利用了一种或者两种假设。后来,Jesse 提出了一种不利用任何显式假设的奇异的半监督分类方法——蕴含限制的最小二乘法(ICLS)^[17]。ICLS 在多维情况下比全监督最小二乘法(LS)分类更加准确,且在一维情况下分类精度不会低于全监督最小二乘法。

在已提出的大量的半监督分类方法中,既有利用聚类假设和流行假设中的一种的方法,也有同时利用两种的方法,甚至不利用任何假设的方法。但是很难在真实的半监督学习任务中决定采用哪种方法或假设,先前研究者们都致力于研究能够提高分类精度的新方法^[18-19],几乎没有把精力用于比较各个方法的分类效果^[20-21]。针对真实的学习任务中选择何种半监督分类方法这一问题,对典型半监督分类方法进行了比较。由于半监督分类方法可以通过数据分布假设来划分种类,因此,比较了几种较典型的应用不同假设的方法。以 LS 为基准,比较了 TSVM、LapRLSC,同时利用两种假设的半监督 Boosting 算法,以及 ICLS 在真实数据集上的分类效果。

2 半监督分类方法

简要介绍了分析研究的典型半监督分类算法,包括利用聚类假设的 TSVM,利用流行正则化的 LapRLSC,同时利用聚类假设和流行假设的半监督 Boosting 算法,以及不利用任何假设的 ICLS。首先给定实

验标注:一组有标记样本 $X_l = \{x_i\}_{i=1}^l$ 和相应的标签 $Y = \{y_i\}_{i=1}^l$,以及无标记样本 $X_u = \{x_j\}_{j=l+1}^n$ ($x_i \in R^d$, $y_i \in \{+1, -1\}$, 无标记样本数量 $u = n - l$)。

2.1 转导支持向量机

聚类假设是两种常见的半监督分类假设中的一种,它假设在同一簇聚类中的样本具有相同的标签,分类边界穿过低密度区域来划分不同的簇。因此,聚类假设也被称作低密度分离假设。TSVM 是利用聚类假设的半监督算法的典型代表。

对于给定的标记样本和未标记样本,TSVM 同时寻找合适的决策边界和未标记样本的标签 $y_1^*, y_2^*, \dots, y_k^*$ 。TSVM 的目标函数通常采用如下形式:

$$\begin{aligned} & \text{Minimize over } (y_1, y_2, \dots, y_u^*, w, b, \\ & \quad \xi_1, \xi_2, \dots, \xi_l, \xi_1^*, \xi_2^*, \dots, \xi_u^*) \\ & \frac{1}{2} \|w\|^2 + C \sum_{i=0}^l \xi_i + C^* \sum_{j=0}^u \xi_j^* \\ & \text{s. t. } \forall_{i=1}^l : y_i [w \cdot x_i + b] \geq 1 - \xi_i \\ & \quad \forall_{j=1}^u : y_j^* [w \cdot x_j^* + b] \geq 1 - \xi_j^* \\ & \quad \forall_{i=1}^l : \xi_i > 0 \\ & \quad \forall_{j=1}^u : \xi_j^* > 0 \end{aligned} \quad (1)$$

其中, C 和 C^* 为正则化参数; ξ_i 和 ξ_j^* 为松弛变量; $C^* \sum_{j=0}^u \xi_j^*$ 为无标签样本 j 在目标函数中的惩罚项。

TSVM 首先通过归纳式 SVM 在训练集上训练得到初始分类器,然后把无标记样本分为正类或负类,再根据目标函数的下降程度,转换无标记样本的类别标签,最后通过迭代的策略解决最优化问题(1)。

2.2 流行正则化

另一种半监督分类学习中常用的假设是流行假设,该假设设想数据在低维服从流行分布。这个数据流行内在的几何结构通常由拉普拉斯图表示,图中的顶点代表样本,图中的边权值代表样本间的相似度。根据流行假设可知,图中相似的节点具有相同的标签。流行正则化^[12,22]是基于流行假设的经典算法,该算法充分利用流行分布的几何结构,并且将数据的几何结构与数据间的相似度约束相结合,作为附加的正则化项添加到目标函数中。

LapRLSC 求解一个带有最小二乘损失函数的最优化问题:

$$\begin{aligned} & \min_{f \in H_k} \frac{1}{l} \sum_{i=1}^l V(x_i, y_i, f) + \gamma_A \|f\|_k^2 + \\ & \quad \frac{\gamma_l}{(u+l)^2} \sum_{i,j=1}^n (f(x_i) - f(y_i))^2 W_{ij} \end{aligned} \quad (2)$$

其中, γ_A 和 γ_l 为正则化参数; $V(x_i, y_i, f)$ 为损失函数,可以是应用于最小二乘法的二乘损失

$(y_i - f(x_i))^2$ 或者是应用于 SVM 的铰链损失 $\max[0, 1 - y_j f(x_i)]$, 从这两方面展开, 流行正则化框架发展为具体的正则化支持向量机 (LapSVM) 和正则化最小二乘法 (LapRLSC)。目标函数中的第三项可以写作 $\frac{\gamma_l}{(u+l)^2} \mathbf{f}^T \mathbf{L} \mathbf{f}$, 其中 $\mathbf{f} = [f(x_1), f(x_2), \dots, f(x_n)]^T$, \mathbf{L} 是由 $\mathbf{L} = \mathbf{D} - \mathbf{W}$ 计算得到的拉普拉斯图, 对角矩阵 \mathbf{D} 由 $D_{ii} = \sum_{j=1}^n W_{ij}$ 构成, 当 K 时, 相当于无标记样本的系数为 0, 此时的优化问题等价于标准的全监督正则化最小二乘法 (RLSC)。

选择最小二乘损失作为损失函数, 因此, 正则化最小二乘法的目标函数可以写为:

$$\min_{\mathbf{f} \in H_K} \frac{1}{l} \sum_{i=1}^l (y_i - f(x_i))^2 + \gamma_A \|\mathbf{f}\|_K^2 + \frac{\gamma_l}{(u+l)^2} \mathbf{f}^T \mathbf{L} \mathbf{f} \quad (3)$$

由申述定理知, $f(x) = \sum_{i=1}^n \alpha_i K(x, x_i)$, $\alpha_i \in R^{C \times 1}$ 。因此, 转换公式为:

$$\argmin_{\alpha \in R^c} \frac{1}{l} (Y - \mathbf{J} \mathbf{K} \alpha)^T (Y - \mathbf{J} \mathbf{K} \alpha) + \gamma_A \alpha^T \mathbf{K} \alpha + \frac{\gamma_l}{(u+l)^2} \alpha^T \mathbf{K} \mathbf{L} \mathbf{K} \alpha \quad (4)$$

其中, $\alpha = [\alpha_1, \alpha_1, \dots, \alpha_n] \in R^{C \times n}$ 为拉格朗日乘子矩阵; \mathbf{K} 为 $n \times n$ 的核矩阵; $\mathbf{J} = \text{diag}(1, \dots, 1, 0, \dots, 0)$ 为对角矩阵。

对目标函数求导, 即可求得最优解。

2.3 半监督 Boosting

SemiBoost 是一种同时利用聚类假设和流行假设, 并且利用 boosting 框架训练分类器的半监督分类算法。用户可以选择一个偏爱的全监督分类器, 然后吸收无标记样本来提升分类器的表现。对于每个无标记样本 x_j , 分别计算对其分类为正类或为负类的置信因子, 其中被当前分类器赋予置信因子最高的无标记样本标签叫做“伪标签”。在循环过程中, 将这些带有伪标签的样本和有标记样本一同加入训练集来训练分类器, 经历一定次数的循环后, 形成最终的分类器。SemiBoost 模型可以定义如下:

$$\argmin_{h(x), \alpha} \sum_{i=1}^l \sum_{j=1}^u S_{i,j} \exp(-2y_i^l (H_j + \alpha h_j)) + C \sum_{i,j=1}^n S_{i,j} \exp(H_i - H_j) \exp(\alpha (h_i - h_j))$$

$$\text{s. t. } h(x_i) = y_i, i = 1, 2, \dots, l \quad (5)$$

其中, $S_{i,j}$ 为样本对 S_i 和 S_j 之间的相似度; C 为有标记样本和无标记样本之间的权重; $H(x) =$

$\sum_{t=1}^T \alpha_t h^{(t)}(x)$ 为前 T 次迭代后形成的分类器, α_t 为组合权重, $h^{(t)}(x)$ 为第 t 次迭代时通过全监督算法 (LS) 学习得到的分类器。

该优化问题可以近似写成:

$$\sum_{i=1}^n (p_i + q_i) (e^{2\alpha} + e^{-2\alpha} - 1) - \sum_{i=1}^n 2\alpha h_i (p_i - q_i) \quad (6)$$

其中, $p_i = \sum_{j=1}^l S_{i,j} e^{-2H_j} \delta(y_j, 1) + \frac{C}{2} \sum_{j=1}^u S_{i,j} e^{H_j - H_i}$ 表示

选取无标记样本 x_i 分类为正标签的置信度; $q_i = \sum_{j=1}^l S_{i,j} e^{2H_j} \delta(y_j, -1) + \frac{C}{2} \sum_{j=1}^u S_{i,j} e^{H_j - H_i}$ 表示将选取的无标记样本 x_i 分为负标签的置信度。当 $x=y$ 时, $\delta(x, y)$ 的值为 1, 否则为 0。

为了最小化该目标函数, 将选取样本 x_i 赋予的最优类别标记为 $z_i = \text{sign}(p_i - q_i)$, 选取样本的权重为 $|p_i - q_i|$, 并且参数 α 的取值应为:

$$\alpha = \frac{1}{4} \ln \frac{\sum_{i=1}^n p_i \delta(h_i, 1) + \sum_{i=1}^n q_i \delta(h_i, -1)}{\sum_{i=1}^n p_i \delta(h_i, -1) + \sum_{i=1}^n q_i \delta(h_i, 1)} \quad (7)$$

初始化 $H(x) = 0$, 每次迭代由全监督分类算法 (LS) 学习得到 $h(x)$, 并且更新分类器 $H(x) = H(x) + \alpha h_i(x)$ 。

2.4 蕴含限制的最小二乘法

不同于以上利用一种或两种数据分布假设的半监督分类算法, ICLS 不利用任何明确的假设, 只利用已经存在于全监督最小二乘分类器中蕴含的假设。ICLS 通过最小化一个常规的全监督分类的损失来得到最优分类器, 其中全监督损失由未标记样本的所有可能标签来定义。ICLS 的目标函数可以写为:

$$\argmin_{\beta \in R^{n \times 1}} \frac{1}{n} \|\mathbf{X} \beta - \mathbf{y}\|^2$$

$$\text{s. t. } \beta \in C_\beta$$

ICLS 实际利用无标记样本来最小化全监督的损失函数。问题(8)的解可以看作是由全监督子集 β 到半监督子集 C_β 的映射, 可以写成如下形式:

$$C_\beta = \{\beta = (\mathbf{X}_e^T \mathbf{X}_e)^{-1} \mathbf{X}_e^T \begin{bmatrix} y \\ y_u \end{bmatrix}, y_u \in [0, 1]^u\} \quad (9)$$

其中, $\mathbf{X}_e = [\mathbf{X}^T, \mathbf{X}_u^T]^T$, 且 β 有关于 y_u 的固定形式, 则在式(8)中, 以 y_u 替代 β 后的最小化方程为:

$$\argmin_{y_u} \frac{1}{n} \left\| (\mathbf{X}_e^T \mathbf{X}_e)^{-1} \mathbf{X}_e^T \begin{bmatrix} y \\ y_u \end{bmatrix} - y \right\|_2^2$$

s. t. $y_u \in [0, 1]^u$

最终, 可得到关于无标记样本的最优解。ICLS 在

多维情况下比全监督最小二乘法分类更加准确,并且在一维情况下分类精度不会低于全监督最小二乘法。

3 实验及结果分析

3.1 数据集和实验设置

数据描述:真实数据集包含了 13 个 UCI 数据集和 6 个基准数据集,具体描述见表 1。

对于真实数据集,同样利用 PCA^[23] 将其映射到 2 维空间,发现一些数据集的数据分布满足聚类假设,例如 Australian、Ionosphere。还有一些数据集满足流行假设,如 Digit1。另外,还有一些数据集同时满足以上两种假设,如 WDBC、USPS。然而,大部分数据集的数据

分布并不明确,比如 Heart、Bupa、House 等。

实验设置:实验中采用高斯核函数,其中高斯核的参数通过所有样本点的平均距离决定。正则化参数 r_A 和 r_I 固定为 1 和 0.1,设置半监督 Boosting 的迭代次数为 12 次。

对于真实数据集的实验,采用十字交叉验证方法^[24],将每个数据集随机分割为 10 等份,然后循环地以一组作为测试集,其余作为训练集。训练集中的有标记样本个数的选取策略与 ICLS^[17] 相同,其中有标记样本是随机选择的,并且个数为 $l = \max\{m + 5, 20\}$, m 为样本特征数。将做 10 次十字交叉验证实验,取其平均值作为结果,实验结果见表 2。

表 1 真实数据集的数据分布

数据集	数据分布	样本个数	特征数	数据集	数据分布	样本个数	特征数
Heart	不明确	270	13	Sonar	不明确	208	60
Australian	聚类	690	14	SPECTF	不明确	267	45
Bupa	不明确	345	6	WDBC	全部	569	30
House	不明确	506	13	Digit1	流行	1 500	241
Vehicle	不明确	435	13	USPS	全部	1 500	241
Haberman	不明确	306	4	COIL2	全部	1 500	241
Ionosphere	聚类	351	34	BCI	不明确	400	117
Parkinsons	不明确	195	22	G241c	聚类	1 500	241
Pima	不明确	768	8	G241n	聚类	1 500	241

表 2 分类错误率比较

数据集	LS	TSVM	LapRLSC	SemiBoost	ICLS
Heart	31±1	28±4	32.8±3	31±3	27±1
Australian	28.5±2	18±5	20.4±4	22±3	22±2
Bupa	41.5±2	39±3	41.5±3	45±4	41.5±2
House	22.2±2	11.6±1	6.9±1	7.3±1	16.3±2
Vehicle	22.5±2	29.5±3	26±4	27±4	18.4±1
Haberman	28±2	27±2	27.6±3	26.5±2	28±2
Ionosphere	28±2	12.4±2	17.5±3	28±4	19±2
Parkinsons	27±2	24.6±2	21±3	21.5±3	24±2
Pima	32±2	37.2±4	36.3±3.5	37±4	30±1
Sonar	44±2	28±3	24.5±2.5	39±4	35±3
SPECTF	44±3	20±1	23.5±4	22.8±2	37±3
WDBC	10±2	10.5±2	10.2±2	12.3±2	8±1
Digit1	42±2	4.6±1	3.3±1	27.4±1	20±1
USPS	42±1	9±1	4.9±1	19±1	20±1
COIL2	38±1	7.8±1	7.6±1	28±3	20±1
BCI	41±3	28.8±3	27.5±2.4	40±3	28±3
G241c	45±1	15.5±2	24.6±3	26.3±3	28±1
G241n	45±2	19±2	23±3	30±4	28±1

3.2 真实数据集的实验结果比较

表 2 列出了真实数据集上的实验结果。根据表 2,可以得出以下结论:
(1)在已知数据集的数据分布,或者能够通过 PCA 降维得到相应数据分布的情况下,基于相应假设的半监督分类方法表现出众。例如,对于满足聚类假设的数据集,TSVM 分类效果最好,对满足流行假设的

数据集,LapRLSC 分类错误率最低。另外,LapRLSC 在同时满足两种假设的数据集上同样有较低的错误率。

(2)当给定数据集不满足任何数据分布假设,并且强调分类安全性时,ICLS 会是明智的选择。原因是 ICLS 分类精度不会低于全监督最小二乘法,ICLS 对于无标记样本的使用不会恶化分类效果。同时,ICLS 在 Heart、Vehicle 和 Pima 数据集上的分类精度是所有半

监督分类算法中最高的,而 LapRLSC 在这些数据集上的分类精度低于全监督 LS,TSVM 和 SemiBoost 同样不能保证分类效果优于全监督算法。

(3)SemiBoost 同时利用流行假设和聚类假设,并采用迭代的 Boosting 算法框架,但是分类效果并没有期望的出色。因此,需要在未来的工作中寻找更有效的算法来结合这些假设,并发挥它们的长处。

3.3 健壮性比较

从真实数据集中选择 5 个数据集,在赋予不同有标记样本个数的情况下比较不同算法的健壮性。对于每个数据集,每次在训练集中随机选取 5,10,20,50,100 个样本赋予标记,并且采用交叉验证的实验设置,每组实验重复 10 次,取其平均值作为结果,实验结果见表 3。

表 3 健壮性比较

算法	数据集	5	10	20	50	100
LapRLSC	Haberman	35.17±6.83	29.50±5.54	27.55±4.33	26.27±1.85	25.11±2.60
	Parkinson	33.06±5.96	27.68±4.26	22.76±3.98	17.98±2.44	15.43±3.01
	House	14.42±5.53	10.91±3.64	7.91±1.18	6.94±0.83	6.80±0.67
	WDBC	18.52±6.88	15.17±3.03	12.46±3.90	9.77±2.36	9.16±1.07
	Australian	31.62±9.24	24.33±3.20	20.37±3.56	17.60±2.69	15.64±1.32
TSVM	Haberman	29.34±4.23	28.16±2.82	27.12±2.12	26.82±0.98	26.50±0.23
	Parkinson	28.57±4.53	26.64±3.65	25.12±1.85	24.50±1.02	24.50±0.30
	House	16.00±7.37	13.68±4.09	11.66±2.08	7.49±1.56	7.09±0.77
	WDBC	18.26±4.32	15.10±3.81	12.50±3.78	8.17±1.26	7.13±0.78
	Australian	28.24±6.68	22.34±5.18	18.24±4.10	16.38±2.15	15.07±0.87
ICLS	Haberman	37.88±5.68	32.83±2.42	28.12±2.20	25.86±0.72	25.71±0.69
	Parkinson	35.81±8.34	31.47±5.86	26.74±2.14	19.64±1.34	14.21±1.40
	House	36.40±6.02	24.93±4.83	16.35±2.47	9.15±1.75	7.27±0.84
	WDBC	21.74±5.27	15.24±3.36	12.52±2.95	7.53±1.76	6.97±1.22
	Australian	38.85±7.26	28.15±4.83	22.34±2.41	16.65±1.62	14.82±1.19
SemiBoost	Haberman	31.24±6.76	28.32±5.15	26.52±3.72	26.34±2.87	26.00±1.42
	Parkinson	28.69±4.53	25.26±5.08	20.82±4.23	19.94±3.05	18.11±2.29
	House	11.14±2.86	9.95±1.25	8.33±0.60	7.42±0.38	7.11±0.40
	WDBC	31.77±6.22	22.74±5.29	17.24±4.71	11.16±2.32	10.65±1.81
	Australian	31.43±6.77	26.17±6.12	24.34±5.18	18.21±3.79	16.27±2.12

根据表 3 可以看出:TSVM 最稳定,健壮性最好。尤其在有标记样本数目较少的情况下,TSVM 是分类精度最高的算法,但其精度并没有随着有标记样本数目的增加而增加。因此,TSVM 适用于给定有标记样本数目有限的情况;ICLS 和 LapRLSC 的分类精度明显地随着有标记样本的个数的改变而改变。当有标记样本数目较少时,无论是 ICLS 还是 LapRLSC 都没有 TSVM 的分类精度高,但它们的分类精度随着有标记样本数目的增长而明显增长。所以,ICLS 和 LapRLSC 适用于给定有标记样本较充裕的情况;SemiBoost 无论是健壮性还是分类精度,表现都相对一般。

3.4 分析讨论

观察以上实验结果,可得到一些发现,并期望给选择哪种半监督分类算法做出一些指导。

(1)在可以明确数据集的数据分布的情况下,利用相应假设的半监督分类算法能保证最好的分类效果。但在现实应用中很难得知数据的内在分布信息。

(2)若对于数据的真实分布没有任何先验知识,将很难判断哪种半监督分类算法比较适合目前的学习

任务。从以上实验结果可知,在有标记样本数目较少的情况下,TSVM 是分类精度最高的算法。因此,TSVM 适用于给定有标记样本数目有限的情况,即使其精度并没有随着有标记样本数目的增加而明显增加。

(3)ICLS 是不利用任何假设的半监督分类算法。研究者们已经证明了在假设不正确或有误差时,无标记样本有可能降低分类精度,而 ICLS 的分类精度却从不低于全监督 LS。因此,若能获取一定量的有标记样本,并强调分类的安全性,尽管 ICLS 相对于全监督算法的精度提升不是那么明显(尤其是在基准数据集上),仍然是最合适的算法。

(4)LapRLSC 在满足流行假设,甚至满足聚类假设的数据集上的分类效果比较令人满意。即使在某些情况下,LapRLSC 的分类精度低于全监督算法,但从总体上看,LapRLSC 的分类效果最好。所以当有标记样本不那么稀缺时,LapRLSC 是一个不错的选择。

(5)尽管 SemiBoost 同时利用流行假设和聚类假设,但在以上的实验中,SemiBoost 并没有令人印象深刻的表现。因此,其他更有效地利用多种假设的策略

仍然值得研究。

4 结束语

大量的半监督分类方法在理论上取得了长足进展,既有利用聚类假设或流行假设其中之一的数据分布假设的方法,也有利用两种数据分布假设的方法,还有不利用任何假设的方法。因此,在真实的半监督学习任务中,采用哪种方法或者假设确实是一个难题。文中比较了利用聚类假设的 TSVM,利用流行假设的 LapRLSC,同时利用以上两种假设的半监督 Boosting 算法,以及不利用任何假设的 ICLS 在真实数据集上的分类效果。

实验结果表明,在已知数据分布的情况下,应该选择利用相应假设的半监督分类算法来保证获得较高的分类精度;若事先不知道样本的数据分布,并且给定的已标记样本数量有限,可以优先选择 TSVM;若具有一定数量的有标记样本,并且强调分类的安全性,不利用任何假设的 ICLS 是比较合适的算法;另外, LapRLSC 也是一个不错的选择。在真实的应用中还存在一些满足多种数据分布的数据集,将在未来的工作中寻找一种将多种假设结合的更有效的算法。

参考文献:

- [1] Fujino A, Ueda N, Nagata M. Adaptive semi-supervised learning on labeled and unlabeled data with different distributions [J]. Knowledge and Information Systems, 2013, 37(1): 129-154.
- [2] Sun S, Hussain Z, Shawe-Taylor J. Manifold-preserving graph reduction for sparse semi-supervised learning[J]. Neurocomputing, 2014, 124(2): 13-21.
- [3] 梁吉业,高嘉伟,常瑜. 半监督学习研究进展[J]. 山西大学学报:自然科学版, 2009, 32(4): 528-534.
- [4] Zhu S P, Huang H Z, Li Y, et al. Probabilistic modeling of damage accumulation for time-dependent fatigue reliability analysis of railway axle steels[J]. Journal of Rail and Rapid Transit, 2015, 229(1): 23-33.
- [5] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th international conference on machine learning. Bled, Slovenia: [s. n.], 1999: 200-209.
- [6] Wang Y, Chen S, Zhou Z H. New semi-supervised classification method based on modified cluster assumption[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(5): 689-702.
- [7] 高滢,刘大有,齐红,等. 一种半监督 K 均值多关系数据聚类算法[J]. 软件学报, 2008, 19(11): 2814-2821.
- [8] 李昆仑,曹铮,曹丽苹,等. 半监督聚类的若干新进展

- [J]. 模式识别与人工智能, 2009, 22(5): 735-742.
- [9] Subramanya A, Talukdar P P. Graph-based semi-supervised learning[C]//Synthesis lectures on artificial intelligence and machine learning. [s. l.]: [s. n.], 2014.
- [10] Ugander J, Backstrom L. Balanced label propagation for partitioning massive graphs[C]//Proceedings of the sixth ACM international conference on web search and data mining. [s. l.]: ACM, 2013: 507-516.
- [11] 肖宇,于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813.
- [12] Belkin M, Niyogi P, Sindhvani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006, 7: 2399-2434.
- [13] Mallapragada P K, Jin R, Jain A K, et al. Semiboost: boosting for semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2000-2014.
- [14] 侯杰,茅耀斌,孙金生. 一种最大化样本可分性半监督 Boosting 算法[J]. 南京理工大学学报:自然科学版, 2014, 38(5): 675-681.
- [15] Freund Y. Experiments with a new boosting algorithm[C]//Thirteenth international conference on machine learning. [s. l.]: [s. n.], 1996: 148-156.
- [16] Chen K, Wang S. Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 129-143.
- [17] Krijthe J H, Loog M. Implicitly constrained semi-supervised least squares classification[C]//International symposium on intelligent data analysis. [s. l.]: Springer International Publishing, 2015: 158-169.
- [18] 李亚娥,汪西莉. 一种自适应的半监督图像分类算法[J]. 计算机技术与发展, 2013, 23(2): 112-114.
- [19] 皋军,王士同,邓赵红. 基于全局和局部保持的半监督支持向量机[J]. 电子学报, 2010, 38(7): 1626-1633.
- [20] Corollary A. A comparative study: globality versus locality for graph construction in discriminant analysis[J]. Journal of Applied Mathematics, 2014, 2014: 1-12.
- [21] Qiao L, Zhang L, Chen S. An empirical study of two typical locality preserving linear discriminant analysis methods[J]. Neurocomputing, 2010, 73(10-12): 1587-1594.
- [22] 柯圣. 基于样本先验信息的正则化型分类器设计研究[D]. 上海: 华东理工大学, 2014.
- [23] Turk M, Pentland A. Eigenfaces for recognition. J Cogn Neurosci[J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.
- [24] Refaailzadeh P, Tang L, Liu H. Cross-validation[M]//Liu L, Zsu M T. Encyclopedia of database systems. New York: Springer, 2009: 532-538.