文章编号:1671-4598(2019)04-0184-06 DOI:10.16526/j.cnki.11—4762/tp.2019.04.042 中图分类号:TP181 文献标识码:A

基于选择性抽样的SVM增量学习算法的泛化性能研究

余 炎[1],徐 婕[1],陈 前[1],杨 艳[2]
(1.湖北大学 计算机与信息工程学院,武汉 430062;
2.武汉晴川学院 计算机科学学院,武汉 430204)

摘要:针对大数据环境中存在很多的冗余和噪声数据,造成存储耗费和学习精度差等问题,为有效地选取代表性样本,同时提高学习精度和降低训练时间,提出了一种基于选择性抽样的SVM增量学习算法,算法采用马氏抽样作为抽样方式,抽样过程中利用决策模型来计算样本间的转移概率,然后通过转移概率来决定是否接受样本作为训练数据,以达到选取代表性样本的目的;并与其他SVM增量学习算法做出比较,实验选取9个基准数据集,采用十倍交叉验证方式选取正则化参数,数值实验结果表明,该算法能在提高学习精度的同时,大幅度的减少抽样与训练总时间和支持向量总个数。

关键词:SVM;增量学习;马氏抽样;转移概率

Research on Generalization Performance of SVM Incremental Learning Algorithm Based on Selective Sampling

Yu Yan[1], Xu Jie[1], Chen Qian[1], Yang Yan[2]
(1. College of Computer and Information Engineering, Hubei University, Wuhan 430062, China;
2. College of Computer Science, Wuhan Qingchuan University, Wuhan 430223, China)

Abstract: For the large data environment, there are many redundancy and noisy data, resulting in storage costs and poor learning accuracy problems, in order to effectively select representative samples, while improving learning accuracy and reduce training time, the thesis presents a selective sampling of SVM incremental learning algorithm, based on Markov sampling as a sampling method. In the sampling process, the decision-making model is used to calculate the transition probability between samples, and then the transition probability is adopted to decide whether to accept the sample as the training data, in order to select the representative sample. Compared with other SVM incremental learning algorithms, the experiment selects 9 benchmark data sets and uses 10-fold cross-validation to select regularization parameters, and the numerical experiments show that the algorithm can greatly reduce the total time of sampling and training and the number of support vectors while improving learning accuracy.

Keywords: SVM; incremental learning; Markov sampling; transition probability

0 引言

基于支持向量机(support vector machine, SVM)的分类算法[1],不仅在解决非线性、小样本、高维模式识别和克服“维数灾难”等问题上中表现出了特有的优势,而且还具有坚实的统计学习理论基础[2-3],简洁的数学模型以及良好的泛化性能。因此,SVM被广泛应用到时间序列预测[4]、回归分析[5]、人脸图像识别等各个领域。尽管SVM理论基础坚实,泛化性能良好,但经典SVM算法是批量式处理,即训练样本一次性被输入到计算机内存中,所以在处理大规模数据时会面临内存限制[6]以及学习效率低等问题。因此具有增量学习功能的数据分类技术应运而生,Syed等人[7]最早提出增量学习,其解决问题的核心在于每一次随机选取算法能够处理的数据量进行训练,留下支持向量,再加入新的训练样本继续训练,依此过程训练学习。近年来,孔锐等人[8]提出一种新的SVM增量学习算法,该算法首先选择可能成为支持向量的边界向量,以达到减少训练的样本数量。Li等人[9]提出基于超平面距离的支持向量机增量学习算法,采用Hyperplane-Distance方法提取样本,选取最有可能成为支持向量的样本构成边缘向量集以提高训练速度。

上述增量学习算法都是基于样本是独立同分布的假设,该假设无论在理论上,还是实际应用中都是非常强的,因现实机器学习[10]中不服从独立同分布的数据很是广泛,所以为非独立同分布的数据更适用于机器学习,减弱独立同分布的假设得到了相关学者的关注,Zou等人[11]证明了具有一致遍历马尔可夫链样本的ERM算法是一致的,且一致遍历马尔可夫链在SVM中也得到了应用,如Xu等人[12]证

收稿日期:2018-10-15; **修回日期**:2018-10-24。
基金项目:国家自然科学基金(61370002,61403132)。
作者简介:余 炎(1992-),男,河南信阳人,硕士研究生,主要从事机器学习方向的研究。
徐 婕(1975-),女,湖北武汉人,硕士生导师,教授,主要从事计算机网络、机器学习方向的研究。

明了 SVM 泛化性能马氏抽样要优于随机抽样。针对样本是独立同分布的假设在实际应用中相对牵强，且独立随机抽样的时效普遍偏低，数据存在非全局性等缺点，提出一种新的 SVM 增量学习算法。该算法利用马氏抽样选取具有一致遍历马尔可夫链性质的训练样本集，研究增量学习的特性，并与基于随机抽样的 SVM 增量学习算法和文献［15］提出的增量学习算法做出比较。分别从分类错误率、支持向量个数和抽样与训练总时间三个方面对比增量学习算法性能，选用基准数据集作为样本数据，经实验表明，基于选择性抽样的 SVM 增量学习算法泛化性能更好。

1　相关知识

针对 SVM 增量学习所涉及到的概念以及一致遍历马尔可夫链等内容，本节将给予介绍以及给出相关的定义。

1.1　支持向量机

基于 SVM 的二分类器，是在给定的空间下，寻找能够分割两类样本且具有最大间隔的超平面。设带有类别标记的输入模式集 $X \subset R^n$ 为二分类数据集，类别标签为 $Y=\{+1,-1\}$，输入集的每一个数据点，都有一个类别标签与之对应即 $X \to Y$，从中取大小为 l 的样本作为原始空间训练集：$S=\{s_1=(x_1,y_1),s_2=(x_2,y_2),\cdots,s_l=(x_l,y_l)\}$，其中 $x_i \in X, y_i \in Y, i=1,2,\cdots,l$。SVM 目标是求解可以分割两类样本点的超平面 $wx+b=0$ 的最优解，将求解问题可以归纳为式（1）二次规划问题：

$$\begin{cases} \min\limits_{w,b,\varepsilon} \dfrac{1}{2}\|w\|^2 + C(\sum\limits_{i}^{l}\varepsilon_i) \\ \text{s.t. } y_i((w \cdot x_i)+b)+\varepsilon_i \geqslant 1 \\ \varepsilon \geqslant 0 \quad i=1,2,\cdots,l \end{cases} \tag{1}$$

其中：C 正则化参数，ε 为松弛变量。

借助 Lagrange 乘子法，转化为对偶问题：

$$\begin{cases} \min\limits_{\alpha} \dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum\limits_{j=1}^{l}\alpha_j \\ \text{s.t. } \sum\limits_{i=1}^{l} y_i \alpha_i = 0 \\ 0 \leqslant \alpha_i \leqslant C, \quad i=1,2,\cdots,l \end{cases} \tag{2}$$

只需求解式（2）即可获取最优分类面，若原始空间中求取的分类面效果不佳，依据泛函理论知识。存在一种满足 Mercer 核条件的核函数：$K(x_i,x_j)=<\Phi(x_i)\cdot\Phi(x_j)>$，可通过线性映射 $\Phi: R^n \to H, x \to \Phi(x)$ 将输入空间映射到 Hilber 空间中，则相应的决策函数为：

$$f(x)=sign\Big[\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\Big]$$

其中非零的拉格朗日乘子（$\alpha_i \neq 0$）对应的样本点称作为支持向量。支持向量个数越少，则表明 SVM 的分类器越稀疏。

1.2　增量学习

当训练样本集（l）规模较大时，由于内存等因素，依靠经典的 SVM 获取分类器是不可行的。因此 Syed 等人[7]提出了增量学习算法，算法核心思想如述：将给定训练集进行数据子集划分，学习过程中依次加入划分的数据子集进行训练。设 D_{TR} 为训练集，将训练集 D_{TR} 划分为 k 个数据子集，数据子集分别记为 $D_{TR}^i, i=1,2,\cdots,k$。则基于随机抽样的 SVM 增量学习算法步骤如下：

步骤 1：训练样本集 D_{TR}^1，留下训练集的支持向量记为 SV_1，加入下一个数据子集 D_{TR}^2，构成新的训练集 $SV_1 \cup D_{TR}^2$；

步骤 2：训练样本集 $SV_1 \cup D_{TR}^2$，留下训练集的支持向量记为 SV_2，加入下一个数据子集 D_{TR}^3，构成新的训练集 $SV_2 \cup D_{TR}^3$；

步骤 3：依次重复以上步骤，直至 $D_{TR}^i (i=k)$。

传统的增量学习算法样本的选择是随机抽样，选取的样本之间不具备关联性。在第 2 节将介绍一种基于选择性抽样的 SVM 增量学习算法。

1.3　一致遍历马尔可夫链

实际应用中很多模型产生的样本在本质上是自然涌现的而非独立同分布，如市场预测，语音识别等，这些数据并不符合机器学习中数据独立同分布的假设。所以通过减弱样本是独立同分布的情形，利用一致遍历马尔可夫链模型进行算法泛化性能研究。如下给出一致遍历马尔可夫链的概念：

定义 (Z,E) 为一个可测空间，则一个随机变量序列 $\{Z_t\}t \geqslant 1$ 以及一系列转移概率测度 $P^n(S \mid z_i), S \in E, z_i \in Z$ 共同构成一个马尔可夫链，假定：

$$P^n(S \mid z_i) := P\{Z_{n+i} \in S \mid Z_j, j<i, Z_i = z_i\}$$

记 $P^n(S \mid z_i)$ 为 n 步转移概率：从初始状态为 z_i 的时刻 i 开始，经过 n 步迭代后状态为 z_{n+i} 属于集合 S 的概率。若转移概率不依赖于在时刻 i 之前的 Z_j 状态集，称具有马尔可夫性质，即：$P^n(S \mid z_i) := P\{Z_{n+i} \in S \mid Z_i = z_i\}$，故马尔可夫链特性：若给定当前状态，则马尔可夫链的将来和过去状态都是独立的。

假设给定测度空间 (Z,E) 上的两个测度为 μ_1 和 μ_2，将测度 μ_1 和 μ_2 的全变差定义为：

$$\|\mu_1 - \mu_2\|_{TV} = \sup_{S \in E} |\mu_1(S) - \mu_2(S)|$$

因此，得出一致遍历马尔可夫的定义：若给定状态集 $\{Z_t\}t \geqslant 1$ 符合条件 $\exists \lambda, 0<\lambda<\infty, \exists \varphi, 0<\varphi<1$，且 $\|P^n(\cdot \mid z) - \varphi(\cdot)\|_{TV} \leqslant \lambda\varphi^n, \forall n \geqslant 1, n \in \mathbb{N}$，其中 $\varphi(\cdot)$ 是 $\{Z_t\}t \geqslant 1$ 的平稳分布，则称 $\{Z_t\}t \geqslant 1$ 为一致遍历马尔可夫链。

2　基于选择性抽样的增量学习算法

基于选择性抽样的 SVM 增量学习算法中利用马氏抽样选取增量样本集，马氏抽样通过定义每一次抽样的转移概率来选择样本数据，构建出具有马尔可夫性质的样本

集。记 D_{TR} 为训练集，D_{TE} 为测试集，T 为增量学习次数，N 为每次增量样本的大小，损失函数[13]定义为 $l(f,z)=(1-f(x)y)_+$。基于选择性抽样的 SVM 增量学习算法步骤如下：

算法 1：SVM 增量学习算法

输入：D_{TR}，D_{TE}，T，N，q。

1）依据增量次数 T 将训练集 D_{TR} 划分为 T 个数据子集，每单个数据子集规模为 D_{TR}/T，数据子集分别记为 D_{TR}^k，$k=1,2,\cdots,T$。

2）分别从每个数据子集中随机选取 N 个样本 $X_{iid}^k:=\{z_i\}_{i=1}^N$，通过 SVM 训练，获取每个数据子集的分类模型和支持向量：f_{iid}^k，SV_{iid}^k，$k=1,2,\cdots,T$。

3）合成新分类模型：$f_1=\dfrac{1}{T-1}\sum_{k=2}^{T}f_{iid}^k$，令 $SV_{mar}^1=SV_{iid}^1$。

4）令 $k=2$。

5）令 $u=1$。

6）从训练集 D_{TR}^k 中随机选取一个样本记为当前样本 z_u。

7）从训练集 D_{TR}^k 中随机选取另一个样本记为候选样本 z_*，计算当前样本 z_u 和候选样本 z_* 之间的比率 P：$P=e^{-l(z_*,f_{k-1})}/e^{-l(z_u,f_{k-1})}$。

8）若 $P=1,y_*y_u=-1$ 或 $P<1$，则依转移概率 P 接受样本 z_*；若 $P=1$、$y_*y_u=1$ 则依转移概率 $P'=min\{1,e^{-y_*f_{k-1}}/e^{-y_uf_{k-1}}\}$ 接受样本 z_*；若有连续 n 个候选样本 z_* 不能被接受，此时依转移概率 $P''=min\{1,qP\}$ 接受样本 z_*，如果 z_* 不能被接受，返回步骤 7），否则令 $u=u+1,z_u=z_*$。

9）如果 $u<N$，返回步骤 7），否则，表示选取到 N 个具有马氏性质的样本 X_{mar}^{k-1}。通过 SVM 训练数据集 $SV_{mar}^{k-1}\cup X_{mar}^{k-1}$，获取分类模型 f_k 和支持向量 SV_{mar}^k，令 $k=k+1$

10）若 $k\leqslant T$，返回步骤 5），否则，获取抽样与训练总时间，支持向量数目，并使用分类模型 f_T 计

算在测试集 D_{TE} 中错分率。

输出：错分率、支持向量个数、抽样与训练总时间

评注 1：算法 1 利用数据子集分类模型的均值来定义起始转移概率，可以避免因初始转移概率的定义而导致算法可能会具有的较大波动性。为快速生成马氏样本集，根据文献［12］的研究，在算法 1 中引进了两个参数 n 和 q，其中 n 为候选样本连续被拒绝的次数，q 为解决当损失函数 $l(f,z)$ 值较小时，在以概率接收候选样本时需要花费大量的时间而引入的常数。

3 数值实验

本章将对实验选取的数据集，实验结果，实验分析做出阐述，为让实验更具有效性与说服力，在实验中，对于同一个数据集，均在数据子集划分、增量次数、每次增量数据量完全相同的情况下进行实验。

在实验结果比较中，记“iid”为基于随机抽样的 SVM 增量学习算法，“Markov”为基于选择性抽样的 SVM 增量学习算法。

3.1 实验参数及数据集

实验选取 Matlab2016a 作为编程软件，在 CPU 为 Inter（R）Core（TM）i7－7500 @2.7 GHz，RAM 为 8 G 的环境中编程（因计算机内存限制，其中数据集 Skin 在 CPU 为 Intel（R）Xeon（R）E5－1603－v4@2.8 GHz，RAM 为 32 G 的环境中实验）。处理高维数据时映射核函数选用高斯径向基函数[14]，算法通过 10 倍交叉验证从候选集［－0.01，－0.1，0，1，10，100，1000，10000］中选取正则化参数 $C=1000$。为更好证明算法的泛化能力，实验分别选取 3 维至 300 维的二分类数据集进行算法的泛化能力研究。实验所选取的 9 个数据集如表 1 所示。

表 1 9 个实验数据集

数据集	维度	训练集	测试集
Skin	3	163371	81686
Magic	10	12680	6340
Letter	16	13333	6667
Image	18	26000	20200
Ijcnn	22	127787	63894
Acoustic	50	73869	24632
Splice	60	20000	43500
Connect	126	40534	16889
GFE	300	18624	9312

3.2 与随机抽样增量学习算法对比

为让实验更具说服力，实验中对于同一个数据集分别进行三次增量实验，即 T 值分别取 10，20，30 次，且每次增量样本会依据算法步骤 1 划分出的数据子集规模而定义较大的值，即 N 值。

实验结果如表 2 所示，其中表的第二列为“数字/数字”，如数据集 Skin 中的 10/8000，表示数据集 Skin 增量 10 次（10 个子集），每次增量的样本数为 8000；20/5000 则表示数据集 Skin 增量 20 次（20 个子集），每次增量的样本数为 5000；为充分的表明基于选择性抽样的 SVM 增量学习算法的泛化能力，实验分别从错误分类率，支持向量个数，抽样与训练总时间三个方面对比基于选择性抽样的 SVM 增量学习算法和基于随机抽样的 SVM 增量学习算法。

由表 3 的实验结果可以看出，基于选择性抽样的 SVM 的增量学习算法无论在 T 与 N 取何值时错误分类率均低于基于随机抽样的 SVM 增量学习算法，且能在保证错分率低的同时，能大幅度减少支持向量个数和抽样与训练总时间。因为基于选择性抽样的 SVM 的增量学习算法中增量样本非随机选取，而是通过计算样本之间的转移概率判断是否接

表 2　数值实验结果

数据集	实验参数	错分率/%		支持向量		抽样与训练总时间	
		Markov	iid	Markov	iid	Markov	iid
Skin	10/8000	5.43	7.06	4376	16939	2645.6	16557.4
	20/5000	5.44	7.00	7012	21437	5352.3	37259.3
	30/3000	5.45	7.18	3087	19234	1158.2	33497.6
Magic	10/1000	28.45	30.71	4616	6410	829.3	1955.8
	20/400	27.81	29.05	2596	5147	309.7	1778.4
	30/200	28.77	30.24	2051	3811	218.0	1104.4
Letter	10/1000	25.34	26.76	3573	6144	415.9	1205.8
	20/500	25.59	27.29	3630	6144	622.9	2005.0
	30/300	25.29	26.72	2471	5446	389.2	3397.4
Image	10/2000	12.81	15.87	3027	8436	507.7	2756.6
	20/1000	12.67	15.34	2482	8344	345.5	4774.4
	30/500	12.59	15.76	1339	6261	134.9	2437.8
Ijcnn	10/8000	6.15	7.90	10371	13677	21991.2	38629.4
	20/4000	5.91	8.25	10323	13766	12963.3	23411.5
	30/2000	5.75	8.28	6501	10249	4813.2	12725.3
Acoustic	10/1000	25.18	27.01	2346	5832	1155.3	5792.8
	20/1000	25.45	26.84	4196	11730	7897.3	65214.2
	30/800	24.20	26.84	4018	13888	9501.9	123350
Splice	10/1200	14.06	14.80	1598	4089	339.3	1735.92
	20/800	13.32	14.48	2431	5520	1021.0	7154.4
	30/400	13.37	14.30	1040	3971	189.1	3291.4
Connect	10/2000	20.63	23.48	4100	9650	2232.9	7887.3
	20/1500	20.41	23.34	6589	14389	8824.9	3641.9
	30/800	20.66	22.95	4007	11483	3932.3	27250.0
GFE	10/1200	16.76	17.68	1987	5002	1456.4	7544.9
	20/600	16.28	17.28	1405	4794	1250.4	1184.4
	30/400	16.07	17.71	1444	4705	1030.1	1329.8

受样本，所以通过马氏抽样选取的样本之间具有关联性，可以很大程度的避开噪声等因素对数据的影响。

为更好地展示实验效果，图 1 给出了基于选择性抽样的 SVM 的增量学习算法和基于随机抽样的 SVM 增量学习算法的实验数据集部分错分率详细对比图、支持向量对比图、抽样与训练总时间对比图。

从图 1 的（a）与（d）中可以看出，基于选择性抽样的 SVM 增量学习算法，在增量样本相同的情况下，随着增量次数的增加，错分率总体呈下降趋势，且错分率逐渐趋于平稳，而基于随机抽样的 SVM 增量学习算法，波动较大。

从图 1 的（b）与（e）中可以看出，无论增量次数 T 和增量样本量 N 取何值，基于选择性抽样的 SVM 增量学习算法比基于随机抽样 SVM 增量学习算法的支持向量数目要少，即分类模型更稀疏。

从图 1 的（c）与（f）中可以看出无论增量次数 T 和增量样本量 N 取何值，基于选择性抽样的 SVM 增量学习算法学习效率有很大程度的提升。

3.3　与文献［15］中算法对比

1）算法对比：自 Syed 等人[7]提出增量学习算法以来，

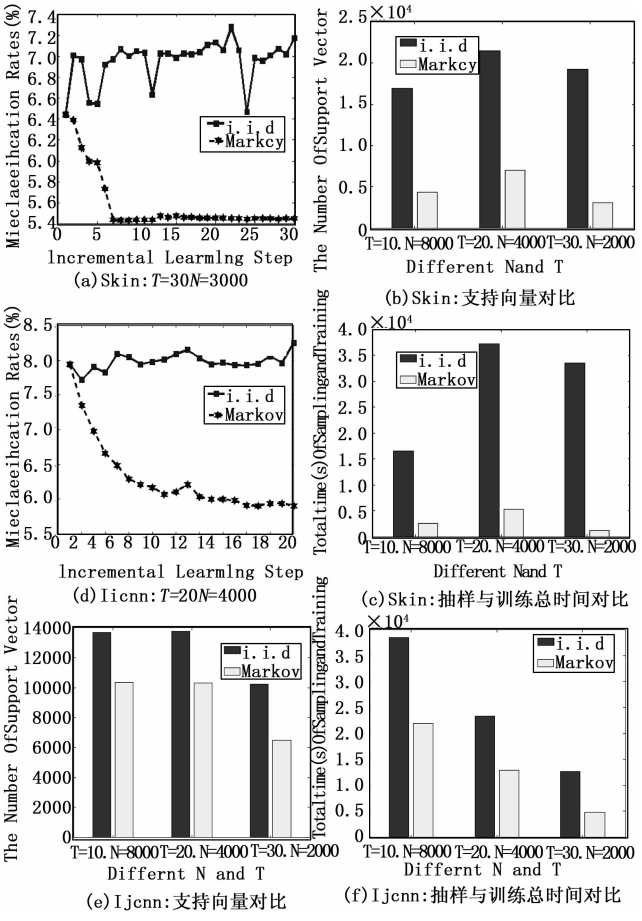

图 1　实验结果详细对比图

以其优异的算法性能得到了许多学者的青睐，同时很多改进的增量学习算法也相继被提出，虽然在算法性能上有一定程度的优化，但基本都是建立在样本是独立同分布的假设情形，本质并没有改变。

Xu 等人[15]提出的增量学习算法，其核心也是利用马氏抽样选取样本进行增量学习（X－ISVM）。基于选择性抽样的 SVM 增量学习算法（M－ISVM）与之最大的区别有以下几点：

（1）X－ISVM 算法在训练集上没有进行子集划分，在整体训练集进行样本选取。M－ISVM 算法则是在每一个数据子集上选取样本。

（2）X－ISVM 算法马氏抽样的初始转移概率是依据第一次随机抽样的分类模型定义，M－ISVM 算法是通过合成 $2 \rightarrow T$ 的数据子集的分类模型来定义马氏抽样的初始转移概率。

（3）文献［15］实验中数据集都以 $T=10$，$N=500$（$T=20$，$N=300$；$T=20$，$N=400$；$T=30$，$N=200$）为基准进行增量学习，增量样本数据量选取数据量较小，不具备说服力；M－ISVM 算法则是根据数据子集规模的大小来定义 N 的值，且 N 值一般定义较大。

2）数值实验与结果：为更好地比较两种算法的泛化能

表 3 T 次($T=10$)增量学习后实验结果对比

数据集	平均错分率/%		平均支持向量		抽样与训练的总时间	
	M－ISVM	X－ISVM	M－ISVM	X－ISVM	M－ISVM	X－ISVM
Skin－500	5.600±0.0006	6.260±0.0067	279.4	467.0	12.57	22.08
Image－500	12.83±0.0018	12.97±0.0033	634.8	635.6	60.07	63.12
Ijcnn－500	5.690±0.0010	5.790±0.0026	483.0	489.8	44.65	48.60
Splice－500	14.02±0.0008	14.74±0.0056	572.8	617.0	114.72	121.45
GFE－500	18.35±0.0108	18.62±0.0126	596.2	620.4	383.48	1209.20

表 4 T 次($T=10$)增量学习后实验结果对比

数据集	平均错分率/%		平均支持向量		抽样与训练的总时间	
	M－ISVM	X－ISVM	M－ISVM	X－ISVM	M－ISVM	X－ISVM
Skin－3000	5.520±0.0007	5.520±0.0011	1709.6	1736.8	791.53	826.21
Image－1500	12.84±0.0024	12.84±0.0025	1934.0	1939.2	1198.50	1939.20
Ijcnn－3000	5.860±0.0020	5.930±0.0030	2937.0	3022.6	4639.60	4713.70
Splice－1000	13.59±0.0015	14.15±0.0023	1226.0	1241.2	930.26	933.30
GFE－1000	17.20±0.0033	18.10±0.0040	1329.4	1485.2	2803.10	3298.80

表 5 T 次($T=20$)增量学习后实验结果对比

数据集	平均错分率/%		平均支持向量		抽样与训练的总时间	
	M－ISVM	X－ISVM	M－ISVM	X－ISVM	M－ISVM	X－ISVM
Skin－300	5.690±0.0037	5.740±0.0048	324.4	446.8	19.45	28.47
Image－300	12.81±0.0015	12.94±0.0024	575.8	619.6	67.01	78.56
Ijcnn－3000	5.560±0.0005	5.690±0.0014	492.4	502.6	45.78	47.19
Splice－300	14.23±0.0042	14.32±0.0047	503.0	534.0	145.54	155.83
GFE－300	18.12±0.0141	18.44±0.0244	544.0	546.2	378.74	382.29

表 6 T 次($T=20$)增量学习后实验结果对比

数据集	平均错分率/%		平均支持向量		抽样与训练的总时间	
	M－ISVM	X－ISVM	M－ISVM	X－ISVM	M－ISVM	X－ISVM
Skin－3000	5.460±0.0001	5.480±0.0003	2279.6	2386.8	2567.40	2672.00
Image－900	12.80±0.0004	12.88±0.0019	1510.6	1584.4	877.70	921.80
Ijcnn－1500	5.660±0.0008	5.760±0.0015	2613.8	2681.6	4794.60	4975.10
Splice－600	13.90±0.0017	14.02±0.0021	974.6	976.0	711.48	766.29
GFE－600	17.70±0.0049	19.11±0.0073	1536.0	1623.1	2055.71	2315.20

力，在基准数据集下，对于每一个数据集分别进行 $T=10$，$N=500$($T=10$ N 依据划分的数据子集规模定义较大值)；$T=20$，$N=300$($T=20$ N 依据数据子集规模定义较大值)的实验，对于每个数据集实验重复 5 次，然后根据每次实验增量最后的分类模型求取五次实验的平均错分率，平均支持向量和 5 次抽样与训练的总时间 (s)。

表 3 为在 $T=10$，$N=500$ 时 X－ISVM 算法和 M－ISVM 算法的平均错分率、方差、平均支持向量和 5 次抽样与训练的总时间的实验数据；表 5 为在 $T=10$ N 依据数据子集规模取值时 X－ISVM 算法和 M－ISVM 算法的平均错分率、方差、平均支持向量和 5 次抽样与训练的总时间的实验数据。

表 5 为在 $T=20$，$N=300$ 时 X－ISVM 算法和 M－ISVM 算法的平均错分率、方差、平均支持向量和 5 次抽样与训练的总时间的实验数据；表 6 为在 $T=20$ N 依据数据子集规模取值时 X－ISVM 算法和 M－ISVM 算法的平均错分率、方差、平均支持向量和 5 次抽样与训练的总时间的实验数据。

表中的第一列为“数据集名－数字”，如“Skin－500”表示从 Skin 数据集中每次增量 500 个训练样本，即算法中 $N=500$。

从表 3～6 中可以看出，M－ISVM 算法，在增量次数

相同的情况下，增量的样本量无论大小，平均错分率，平均支持向量，抽样与训练总时间表现都优于 X－ISVM 算法，且方差更低，说明算法稳定性好。因为 M－ISVM 算法在马氏抽样起始转移概率的定义上利用了 $2\rightarrow T$ 的数据子集的分类模型，而 X－ISVM 算法只利用了第一次随机抽样的分类模型，在每次增量的数据上，M－ISVM 算法分别从每一次数据子集中选取，而 X－ISVM 则在整体训练集中选取，所以 M－ISVM 算法能更好地兼顾全局性，很大程度上避免实验结果的偶然性。实验结果表明，M－ISVM 算法的泛化性能优于 X－ISVM 算法。

4　结束语

传统的增量学习都是建立在样本是独立同分的假设情形下，样本的选取都是基于独立随机抽样，这种假设并不能完全符合实际环境中样本的分布情况。基于选择性抽样的 SVM 增量学习算法，通过减弱样本是独立同分布的假设情形，利用马氏抽样方式选取具有一致遍历马尔可夫链性质的样本进行增量学习，文章中与基于随机抽样的 SVM 增量学习算法和文献［15］提出的算法做出比较。实验结果表明，基于选择性抽样的 SVM 增量学习算法在 SVM 分类问题上泛化性能更好。

参考文献：

［1］Vapnik V N. Statistical learning theory［Z］. New York，NY，USA：Wiley，1998.

［2］Sain S. The Nature of Statistical Learning Theory［J］. IEEE Transactions on Neural Networks，2002，38（4）：409－409.

［3］张学工. 关于统计学习理论与支持向量机［J］. 自动化学报，2000，26（1）：32－42.

［4］Müller K R，Smola A J，Ratsch G，et al. Predicting time series with support vector machines［A］. International Conference on Artificial Neural Networks［C］. Springer Berlin Heidelberg，1997：999－1004.

［5］Rakotomamonjy A. Analysis of SVM regression bounds for variable ranking［J］. Neurocomputing，2007，70（7）：1489－1501.

［6］刘江华，程君实，陈佳品. 支持向量机训练算法综述［J］. 信息与控制，2002，31（1）：45－50.

［7］Syed N A，Liu H，Sung K K. Handling concept drifts in incremental learning with support vector machines［A］. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining［C］. ACM，1999：317－321.

［8］孔　锐，张　冰. 一种快速支持向量机增量学习算法［J］. 控制与决策，2005，20（10）：1129－1132.

［9］Li C，Liu K，Wang H. The incremental learning algorithm with support vector machine based on hyperplane－distance［M］. Kluwer Academic Publishers，2011.

［10］周志华，王　珏. 机器学习及其应用［M］. 清华大学出版社，2007.

［11］Zou B，Xu Z B，Xu J. Generalization bounds of ERM algorithm with Markov chain samples［J］. Acta Mathematicae Applicatae Sinica，2014，30（1）：223－238.

［12］Jie X，Yuan Y T，Zou B，et al. The Generalization Ability of Online SVM Classification Based on Markov Sampling［J］. IEEE Transactions on Cybernetics，2015，45（6）：1169.

［13］Evgeniou T，Pontil M，Poggio T. Regularization Networks and Support Vector Machines［J］. Advances in Computational Mathematics，2000，13（1）：1－50.

［14］Zanaty E A，Afifi A. Support vector machines（SVMs）with universal Kernels［J］. Applied Artificial Intelligence，2011，25（7）：575－589.

［15］Xu J，Xu C，Zou B，et al. New incremental learning algorithm with support vector machines［J］. IEEE Transactions on Systems Man & Cybernetics Systems，2018，（99）：1－12.

（上接第 183 页）

宽度。再基于 VC＋＋平台，开发出车辆痕迹信息管理系统软件。根据该软件查询符合该条件的车辆轮胎规格信息，进而来寻找车辆本身的配置参数、轮胎与对应车型的相关信息以及等轮胎痕迹的特征等内容。

4）利用 Matlab 软件建立轮胎痕迹花纹图像匹配系统，对事故嫌疑车辆轮胎痕迹与数据库轮胎痕迹花纹图像匹配，找出最佳匹配车辆，最后，锁定嫌疑车辆。

参考文献：

［1］杜晓炎，杜心全，李英娟. 道路交通事故现场处理教程［M］. 北京：中国人民公安大学出版社，2005.

［2］徐毅刚. 道路交通事故处理新论［M］. 济南：山东人民出版社，2005.

［3］周奇智. 交通事故现场轮胎痕迹的研究［D］. 西安：长安大学，2011.

［4］宋玲华. 交通肇事逃逸车辆轮胎印痕研究［J］. 江苏公安专科学校学报，2001，15（5）：132－135.

［5］Moser A，Stcffan H. Automatic optimization of pre－impact parameters using post impositions and rest positions［A］. Infats Proceedings of the 3rd International Forum of Automotive Traffic Safety［C］. 1998.

［6］牛学军. 道路交通事故现场勘测［M］. 北京：中国人民公安大学，2007：120－126.

［7］闫松申. 道路交通事故现场轮胎痕迹智能鉴别系统的研究［D］. 吉林：吉林大学，2003.

［8］周　博，谢东来，张宪海. MATLAB 科学计算［M］. 北京：机械工业出版社，2010.