第36卷第11期 东北大学学报(自然科学版) Vol. 36, No. 11
2015年11月 Journal of Northeastern University(Natural Science) Nov. 2015

doi: 10.3969/j.issn.1005-3026.2015.11.005

基于抽样方法的不确定极限学习机

赵相国, 毕 鑫, 张 祯, 喻 鑫
(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 实际应用中的大量数据具有不确定属性,而传统的挖掘算法无法直接应用在不确定数据集上. 针对不确定数据的分类问题,提出一种基于抽样方法的不确定极限学习机. 该算法通过抽样的方法,对不确定数据集中样本的抽样实例进行学习和分类,得到该不确定样本的所属类别的概率,从而实现了传统极限学习机分类算法对不确定数据的分类,并极大降低了不确定对象实例的枚举代价. 实验结果表明,该算法在不确定数据的分类问题中具有较好的有效性和高效性.

关 键 词: 极限学习机;不确定;抽样;分类
中图分类号: TP 311.13 **文献标志码**: A **文章编号**: 1005-3026(2015)11-1539-04

Sampling Based Uncertain Extreme Learning Machine

ZHAO Xiang-guo, *BI Xin*, *ZHANG Zhen*, *YU Xin*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHAO Xiang-guo, E-mail: zhaoxiangguo@mail.neu.edu.cn)

Abstract: Large amounts of data in real-world applications have inherent uncertainty. Traditional learning algorithms cannot be applied directly onto uncertain datasets. Aiming at classification problems over uncertain data, a sampling based uncertain ELM(extreme learning machine) was proposed. Instances were first sampled out of uncertain objects, and then learnt with uncertain ELM. The uncertain objects would be assigned to their classes respectively according to the probabilities aggregation method. The classification was realized by the proposed algorithm in this paper over uncertain data avoiding the enumeration of instances. The experimental results indicated the efficiency and effectiveness of our algorithm.

Key words: ELM(extreme learning machine); uncertain; sampling; classification

极限学习机(extreme learning machine, ELM)因其良好的泛化性能和极快的学习速度,在分类和回归问题中得到了广泛的应用. 然而,由于数据生成和采集过程中的测量误差、噪声、数据缺失等原因,数据集中的训练样本具有不确定性. 包括极限学习机在内的传统分类算法无法直接对不确定数据进行学习,在分类过程中考虑数据的不确定属性是有待解决的问题之一. 因此,本文针对不确定数据分类问题,提出了基于蒙特卡洛抽样的不确定极限学习机(Monte Carlo based extreme learning machine, MC-ELM).

收稿日期: 2014-12-31
基金项目: 国家自然科学基金资助项目(61272181,61202087).
作者简介: 赵相国(1973-),男,吉林通化人,东北大学副教授.

1 问题定义

描述一个不确定对象通常使用离散模型或者连续模型,其中,离散模型使用带有概率值的离散实例集合进行描述,连续模型使用概率密度函数进行描述. 本文基于离散模型对不确定数据的分类问题进行分析. 首先给出离散模型的正式定义,如定义1所描述.

定义1 给定一组不确定对象 $U=\{U_1,\cdots,U_n\}$,其中每个不确定对象 U_i 由一组不确定对象实例组成,即 $U_i=\{u_i^1,\cdots,u_i^m\}$,则有每个不确定对象之间相互独立,每个不确定对象的所有实例

都相互独立,每个不确定对象的所有实例的存在概率之和为1.

在给定不确定数据模型的基础上,本文给出不确定数据分类的问题定义,如定义2所描述.

定义2 给定一个不确定数据集$[U,c]\in X\times C$,其中U是不确定对象,c是不确定对象对应的类别标签,X和C分别是数据和类标签所属的空间.不确定数据分类就是要训练出一个由X到C的映射φ,使得不确定对象U_i的每个实例u_i^j都以一定的概率属于一个类c_i,并且该不确定对象U_i属于类别c,如果属于类别c的U_i的所有实例的分类概率和最大.

最基本的方法是枚举不确定数据集中每个不确定对象的所有实例,使用某种学习算法进行训练之后,计算每个样本的所有实例的分类结果及其概率,并将每个不确定对象的实例中属于同一个类别的实例的分类概率求和,概率和最大的类别即为该不确定对象的所属类别.该方法能够保证不确定数据分类的准确性,但是在现实应用中,枚举所有可能世界的代价太大,几乎是不可操作的,再将得到的庞大的枚举实例训练集进行训练不具有实际意义.

2 极限学习机

极限学习机[1-3]是一种广义单隐层前馈后传神经网络.极限学习机首先随机生成输入权重和偏置,并通过矩阵运算直接计算出输出权重,形成分类器模型.相比传统的前馈后传神经网络,极限学习机的训练速度极快,泛化性能更好,实现更容易.此外,极限学习机和支持向量机(support vector machine,SVM)[4]优化角度是一致的[5].

给定N个训练样本$(x_i,t_i)\in\mathbf{R}^{n\times m}$,极限学习机的输出为

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i G(\boldsymbol{\omega}_i\cdot\boldsymbol{x}+b_i)=\boldsymbol{\beta}h(\boldsymbol{x}). \quad (1)$$

其中:L是隐藏层节点数;$\boldsymbol{\beta}$是输出层权重;$\boldsymbol{\omega}$是输入层权重;$\boldsymbol{b}$是偏置;$G(\boldsymbol{x})$是隐藏层使用的激励函数.$\boldsymbol{H}$是隐藏层映射空间,定义为

$$\boldsymbol{H}=\begin{bmatrix} G(\boldsymbol{\omega}_1\cdot\boldsymbol{x}_1+b_1) & \cdots & G(\boldsymbol{\omega}_L\cdot\boldsymbol{x}_1+b_L) \\ \vdots & & \vdots \\ G(\boldsymbol{\omega}_1\cdot\boldsymbol{x}_N+b_1) & \cdots & G(\boldsymbol{\omega}_L\cdot\boldsymbol{x}_N+b_L)\end{bmatrix}_{N\times L}. \quad (2)$$

输出权重矩阵$\boldsymbol{\beta}$的计算公式为

$$\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{T}. \quad (3)$$

其中:$\boldsymbol{H}^{\dagger}$是特征映射矩阵$\boldsymbol{H}$的广义逆矩阵;$\boldsymbol{T}$是学习样本的类标签矩阵.

极限学习机的计算过程是首先随机生成输入权重$\boldsymbol{\omega}$和$\boldsymbol{b}$,使用随机生成的参数和训练数据$\boldsymbol{x}$利用公式(2)计算矩阵$\boldsymbol{H}$,再使用公式(3)计算输出权重矩阵$\boldsymbol{\beta}$.具体如算法1所描述.

算法1 ELM

输入:训练数据$\boldsymbol{x}$.

输出:输出权重矩阵$\boldsymbol{\beta}$.

01. 随机生成输入权重$\boldsymbol{\omega}$和$\boldsymbol{b}$;

02. 计算$\boldsymbol{H}$矩阵;

03. 计算输出权重矩阵$\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{T}$.

极限学习机的输出函数为

$$f(\boldsymbol{x})=h(\boldsymbol{x})\boldsymbol{\beta}=h(\boldsymbol{x})\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\boldsymbol{C}}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T}. \quad (4)$$

不失一般性,训练数据集中的样本个数要远远大于隐藏层节点数,因此,为了降低计算代价,极限学习机的输出函数可以改写为

$$f(\boldsymbol{x})=h(\boldsymbol{x})\boldsymbol{\beta}=h(\boldsymbol{x})\left(\frac{\boldsymbol{I}}{\boldsymbol{C}}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T}. \quad (5)$$

3 不确定极限学习机

文献[6]提出了一种不确定极限学习机,在枚举所有不确定对象的所有实例后,利用实例的存在概率判断不确定对象对应所属类别的分类概率,从而实现对不确定数据的分类.

然而,枚举不确定数据集中每个不确定对象的所有实例的枚举代价呈指数级增长,尤其是当不确定对象可能的实例很多的时候,枚举代价是巨大的和不可行的.

为此,本节提出一种基于蒙特卡洛方法的不确定极限学习机MC-ELM.该算法使用蒙特卡洛方法对不确定数据集中的不确定对象进行抽样,在对抽样得到的抽样样本进行训练后,计算抽样估计量,并以此作为不确定分类概率.MC-ELM不使用不确定对象的实例的存在概率作为分类概率,而是使用抽样方法的抽样估计量.

3.1 蒙特卡洛抽样和估计量

蒙特卡洛方法是一种随机抽样方法,在特定概率分布的概率模型中,利用抽样实验的方法,产生该已知概率分布的随机变量.基于本文所采用的不确定概率模型,首先给出该随机变量的标记规则.

假定一个不确定对象U_i实际属于类别c_s,u_i^j是U_i的第j个抽样实例,抽样标记$\varepsilon_{u_i^j}^{c_t}$的取值规则为

$$\varepsilon_{u_i^j}^{c_t}=\begin{cases}0, & c_s\neq c_t;\\ 1, & c_s=c_t.\end{cases} \quad (6)$$

其中，c_t 是分类器输出的类别标签．对于每个不确定对象的抽样实例，如果该实例被划分的类别与其不确定对象所属的类别一致，则抽样标记为 1，否则为 0．

不确定对象 U_i 对于类别 c_s 的抽样估计为

$$\mathrm{Pr}_{U_i}^{c_s} \approx \tilde{\mathrm{Pr}}_{U_i}^{c_s} = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_{u_i^j}^{c_s}. \tag{7}$$

估计量 $\mathrm{Pr}_{U_i}^{c_s}$是无偏估计，即 $E(\tilde{\mathrm{Pr}}_{U_i}^{c_s}) = \mathrm{Pr}_{U_i}^{c_s}$，估计偏差为

$$\mathrm{Var}(\tilde{\mathrm{Pr}}_{U_i}^{c_s}) = \frac{1}{n}\mathrm{Pr}_{U_i}^{c_s}(1-\mathrm{Pr}_{U_i}^{c_s}) \approx \frac{1}{n}\tilde{\mathrm{Pr}}_{U_i}^{c_s}(1-\tilde{\mathrm{Pr}}_{U_i}^{c_s}). \tag{8}$$

3.2　基于蒙特卡洛的不确定极限学习机

本节提出一种基于蒙特卡洛的不确定极限学习机（MC－ELM）．与直接对枚举实例进行训练的分类算法不同，MC－ELM 利用上一节介绍的蒙特卡洛方法的抽样估计量对不确定对象的分类概率进行估计．根据蒙特卡洛抽样方法及其估计量的定义，MC－ELM 算法的具体流程如算法 2 所描述．

算法 2　MC－ELM

输入：训练数据集

输出：分类结果

01．ForEach $U_i \in D_{\mathrm{train}}$ DO

02．　对 U_i 进行抽样；

03．　将抽样实例放入 $D_{\mathrm{train}}^{\mathrm{sampled}}$；

04．End ForEach

05．使用 $D_{\mathrm{train}}^{\mathrm{sampled}}$ 计算 $\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T}$；

06．ForEach $U_i \in D_{\mathrm{test}}$ DO

07．　对 U_i 进行抽样；

08．　将抽样实例放入 $D_{\mathrm{test}}^{\mathrm{sampled}}$；

09．End ForEach

10．计算输出 $\boldsymbol{O} = \boldsymbol{H\beta}$；

11．ForEach $U_i \in D_{\mathrm{test}}^{\mathrm{sampled}}$ DO

12．　ForEach $u_i^j \in S_{U_i}$ DO

13．　　从 $\boldsymbol{O}$ 中获取 u_i^j 的类标签 c_{out}；

14．　　对应估计量 $\mathrm{Pr}_{U_i}^{c_{\mathrm{out}}} = \mathrm{Pr}_{U_i}^{c_{\mathrm{out}}} + \frac{1}{n}$；

15．　End ForEach

16．　$c = \mathrm{argmax}_{c_{\mathrm{out}} \in C}\mathrm{Pr}_{U_i}^{c_{\mathrm{out}}}$ 为 U_i 的类标签；

17．End ForEach

首先初始化两个空数据集 $D_{\mathrm{train}}^{\mathrm{sampled}}$ 和 $D_{\mathrm{test}}^{\mathrm{sampled}}$．第 1～4 行对原始训练数据集中的每个不确定对象进行抽样，并将抽样实例存入数据集 $D_{\mathrm{train}}^{\mathrm{sampled}}$，第 5 行根据传统 ELM 的计算理论，使用该数据集中的数据计算分类器的 $\boldsymbol{\beta}$ 参数．第 6～9 行对原始测试数据集中的每个不确定对象进行抽样，并将所有抽样实例存入数据集 $D_{\mathrm{test}}^{\mathrm{sampled}}$．第 10 行根据传统 ELM 的计算理论和测试抽样实例计算输出矩阵 $\boldsymbol{O}$．第 11～17 行判断每个不确定对象的所属类别，其中第 12～15 行计算每个抽样实例对应所有类别的抽样估计量标记，第 16 行将不确定对象划分给抽样估计量总和最大的类别．

4　实验分析

本文实验使用一台配置英特尔酷睿 i7－3667U 3.4 GHz 处理器和 8GB DDR3 内存的商用台式机，ELM 相关算法使用 Matlab R2009a 编写和实现．

实验数据集采用 UCI Machine Learning Respository[7] 中的真实数据集作为原始数据集，并根据一定的概率分布将原始数据集转换成不确定数据集．具体做法是，将原始确定数据集中每个真实的样本作为不确定数据集中一个不确定对象的中心，然后利用概率分布为每个不确定对象随机生成多个实例．原始数据集的信息如表 1 所示．

表 1　数据集信息

Table 1　Brief of datasets

数据集	样本数	类别数
Vowel	1 348	11
Diabete	1 142	2
Credit	1 060	2

由于基于抽样的不确定极限学习机 MC－ELM 的训练时间和准确率都与抽样程度直接相关，因此本节所给出的实验结论都没有与现有算法进行直接对比，而是单独给出与抽样率相关的实验结果变化情况．

MC－ELM 的训练时间随抽样率变化的实验结果图如图 1 所示．由图中可以看出，随着抽样率的增大，抽样实例的个数增加，因此训练时间也相应地增长．

估计量的相对误差能够反映实际学习结果 R 与估计结果 $\tilde{R}$ 的相对差值．本文实验的相对误差使用式（9）计算：

$$\delta = \frac{|R-\tilde{R}|}{R}. \tag{9}$$

MC－ELM 算法的相对误差随着抽样率增大而变化的趋势如图 2 所示．

从图中可以看出，随着抽样率的增加，估计量的相对误差在减小，并逐渐收敛于 0．这一实验结

果表明该估计量具有较高的估计能力.

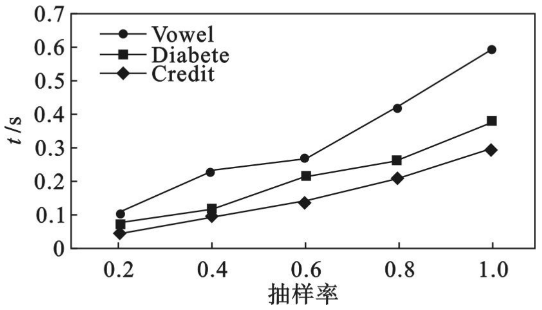

图 1　MC - ELM 的训练时间
Fig. 1　Training time of MC-ELM

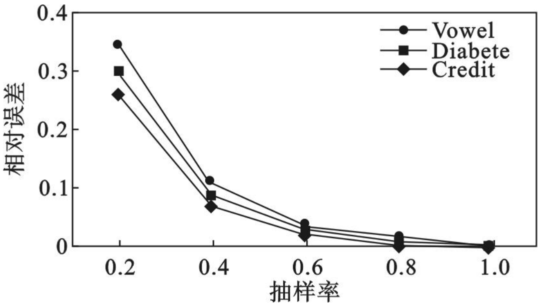

图 2　MC - ELM 的相对误差
Fig. 2　Relative error of MC-ELM

准确率是衡量分类算法的重要指标之一. 本组实验测得 MC - ELM 算法随抽样率增加的测试准确率的变化趋势如图 3 所示.

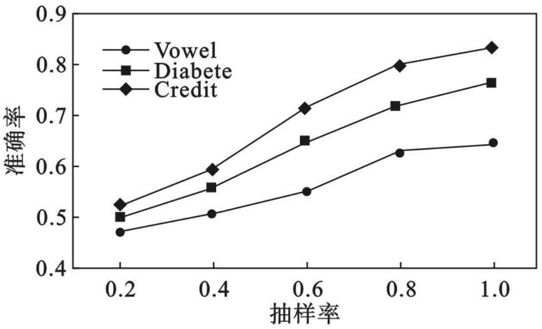

图 3　MC - ELM 的相对误差
Fig. 3　Testing accuracy of MC-ELM

由于抽样估计量具有相对误差,基于抽样的不确定极限学习机的准确率会小于基于存在概率的不确定极限学习机. 当抽样率上升时,分类准确率也随之上升;当抽样率为 1 时,二者理论上相等.

5　结　　语

本文针对不确定数据的分类问题,提出一种基于蒙特卡洛抽样的不确定极限学习机 MC - ELM. 该算法通过对不确定对象进行抽样,通过基于极限学习机计算理论对抽样实例进行学习,并使用抽样估计量代替存在概率计算不确定对象的分类概率. MC - ELM 算法能够有效避免枚举不确定对象实例的代价高等问题. 实验结果表明,本文提出的 MC - ELM 算法在不确定数据分类问题中,具有有效性和高效性.

参考文献:

[1] Huang G B, Zhou H M, Ding X J, et al. Extreme learning machine for regression and multiclass classification[J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 2012, 42(2):513 - 529.

[2] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: a new learning scheme of feedforward neural networks[C]// Proceedings of the 2004 IEEE International Joint Conference on Neural Networks. Budapest, Hungary, 2004:985 - 990.

[3] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. *Neurocomputing*, 2006, 70(1/2/3):489 - 501.

[4] Cortes C, Vapnik V. Support vector networks[J]. *Machine Learning*, 1995, 20(3):273 - 297.

[5] Huang G B, Ding X J, Zhou H M. Optimization method based extreme learning machine for classification[J]. *Neurocomputing*, 2010, 74(1/2/3): 155 - 163.

[6] Sun Y J, Yuan Y, Wang G R. Extreme learning machine for classification over uncertain data[J]. *Neurocomputing*, 2014, 128(27):500 - 506.

[7] Blake C L, Merz C J. UCI repository of machine learning databases[D]. California: University of California, 1998.