针对小规模数据集的多模型融合算法研究

李春生,曹 琦,于 澍
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:目前,对小规模数据集进行预测时,主要使用传统机器学习算法,但传统单一模型预测效果不能达到预期准确率,且无法兼顾多项评价指标。因此,文中以小规模数据集为研究对象,融合决策树、逻辑回归、支持向量机三类模型,提出了一种多模型融合算法,并分析了其在小规模数据集上的应用效果。首先,简述了决策树、逻辑回归和支持向量机的算法原理;其次,使用决策树、逻辑回归和支持向量机作为基学习器并完成单独训练,将各模型输出结果用于下一阶段模型输入,同时使用最大似然估计迭代优化参数,从而完成多模型融合过程;最后,对数据集进行分析和处理,通过实验与单一模型进行指标对比。实验结果表明,多模型融合算法在预测精确率、召回率、准确率等方面有明显提升。

关键词:数据挖掘;机器学习;逻辑回归;决策树;模型融合
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2020)02-0063-04
doi:10.3969/j.issn.1673-629X.2020.02.013

Research on Multi-model Fusion Algorithm for Small Scale Data Sets

LI Chun-sheng,CAO Qi,YU Shu
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:At present,traditional machine learning algorithms are mainly used in the prediction of small-scale data sets,but the traditional single model cannot reach the expected accuracy in prediction effect and cannot take into account multiple evaluation indexes. Therefore,taking the small-scale data sets as research objects and integrating decision tree,logistic regression and support vector machine,we propose a multi-model fusion algorithm and analyze its application effect on small-scale data sets. Firstly,the algorithm principle of decision tree,logistic regression and support vector machine is briefly described. Secondly,decision tree,logistic regression and support vector machine are used as the base learner and the individual training is completed. The output results of each model are used for the model input in the next stage,and the maximum likelihood estimation is used for iterative optimization parameters to complete the multi-model fusion process. Finally,the data sets are analyzed and processed,and the indicators are compared with the single model through experiments which show that this algorithm has a significant improvement in prediction precision,recall rate and accuracy.

Key words:data mining;machine learning;logistic regression;decision tree;model fusion

0 引 言

机器学习作为人工智能的重要研究内容,经过半个世纪的发展,现今已和模式识别、数据挖掘、统计学习、计算机视觉、自然语言处理等多个领域相互影响、交织发展[1]。

集成学习目前是机器学习领域中的一种研究方向。使用弱学习器通过多模型融合的思想可以极大提高准确率。当前集成学习(Bagging)主要使用弱学习器,且为同类模型,例如随机森林使用多棵深度较浅的决策树,在构建 Bagging 集成的基础上将决策树作为基学习器[2],最终进行投票获得最终结果。文中尝试使用多类强学习器进行模型融合,并与单一强学习器进行指标对比。

1 相关研究

1.1 决策树模型

决策树是一个有监督的机器学习算法,常用于分类预测等诸多领域,由于其高效性、误差小的优点,在

收稿日期:2019-02-26 **修回日期**:2019-06-27 **网络出版时间**:2019-11-07
基金项目:国家自然科学面上项目(51774090);黑龙江省自然科学基金面上项目(F2015020);黑龙江省教育科研专项引导性创新基金项目(2017YDL-12);黑龙江省教育规划重大课题(GJ20170006)
作者简介:李春生(1960-),男,博士,教授,博导,研究方向为数据挖掘与智能系统、软件集成技术、图像处理与模式识别、智能仪器与计算机控制系统;曹 琦(1993-),男,硕士研究生,研究方向为机器学习、数据挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0910.026.html

分类问题中得到了广泛的应用。在决策树中,内部分支节点表示一个条件属性,而叶子节点表示一种决策属性或分类结果[3-5]。决策树是一个预测模型,其叶节点代表最终样本分类,各属性划分代表分类规则。

由于文中解决二分类问题,选用当前较为流行的 C4.5 算法作为其中一种基类模型,C4.5 算法是由 J. Ross Quinlan 开发并且用于决策树的算法[6]。C4.5 算法流程与 ID3 类似,相比 ID3,将信息增益改为信息增益比,选择信息增益比大的特征当作决策树的节点并不停递归构建决策树,同时设置阈值避免过拟合。主要公式如下:

数据集 S 的信息熵:

$$H(S) = -\sum_{k=1}^{K} \frac{|B_k|}{N} \log_2 \frac{|B_k|}{N}$$

特征 F 对于数据集 S 的条件信息熵:

$$H(S/F) = -\sum_{i=1}^{n} \frac{N_i}{N} \times H(S_{ik})$$

特征 F 的信息增益:

$$\text{Gain}(S,F) = H(S) - H(S/F)$$

特征 F 对数据集 S 的分裂信息:

$$H_F(S) = -\sum_{i=1}^{n} \frac{N_i}{N} \log_2 \frac{N_i}{N}$$

特征 F 对数据集 S 的信息增益比[7]:

$$\text{GainRatio}(S,F) = \frac{\text{Gain}(S,F)}{H_F(S)}$$

1.2 逻辑回归模型

逻辑回归(logistic regression)是一种可以用来分类的常用统计分析方法,并且可以得到概率型的预测结果,属于一种概率型非线性回归[8-10]。逻辑回归是经典的分类模型,它将模型拆分为线性部分和激活函数,主要公式如下:

假设 x 为输入变量, $\boldsymbol{W}$ 为权重矩阵, B 为偏置, A 为线性部分输出,则线性部分函数为:

$$A = \boldsymbol{W}x + B$$

激活函数使用 sigmoid 函数,将线性部分输出 A 当作 sigmoid 函数输入值, y 为预测结果:

$$y = \frac{1}{1 + e^{-A}}$$

使用交叉熵损失函数,其中 y^i 代表第 i 个样本的预测值,对应的 $\bar{y^i}$ 代表样本的正确输出:

$$\text{Loss} = -\frac{1}{m}\sum_{i=1}^{m} [\bar{y^i}\log(y^i) + (1 - \bar{y^i})\log(1 - y^i)]$$

通过使用梯度下降或者 mini-Batch 梯度下降等算法完成对模型损失函数的迭代,最终给出权重 W 和偏置 B。

1.3 SVM 模型

文献[11-15]指出支持向量机(support vector machines,SVM)是一种二分类模型,它的基本模型是定义在特征空间上的间隔最大的线性分类器,间隔最大使它有别于感知机。文中选用线性可分支持向量机,通过核函数与软间隔最大化,学习得到分类决策函数:

$$f(x) = \text{sign}(\sum_{i=1}^{N} a_i^* y_i K(x,x_i) + b^*)$$

其中 $K(x,x_i)$ 为正定核函数,使用序列最小最优化(sequential minimal optimization,SMO)算法实现支持向量机的优化过程。SMO 算法要解决的是凸二次规划的对偶问题:

$$\min_{\partial} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \partial_i\partial_j y_i y_j K(x_i,x_j) - \sum_{i=1}^{N} \partial_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \partial_i y_i = 0,\ 0 \leqslant \partial_i \leqslant C, i = 1,2,\cdots,N$$

SMO 基本思路为选择两个变量,固定其他变量,针对这两个变量构建一个二次规划问题,这时子问题可以极大提高算法的运算速度。SMO 算法将原问题不断分解为子问题并对子问题进行求解,进而达到求解原问题的目的。

2 多模型融合算法

2.1 基本思想

多模型融合算法思想与 Bagging 集成学习算法思想类似,对比 Bagging 集成学习将弱学习器当作基学习器,使用平均投票得出最终结果的方式。文中提出的多模型融合算法使用强学习器决策树、逻辑回归、SVM 作为基学习器,并将基学习器输出当作下一阶段的输入,加入权重矩阵并使用最大似然估计迭代优化参数,计算出基学习器模型的输出权重参数,从而完成多模型融合过程。

2.2 算法描述

多模型融合算法共分为两部分:基学习器训练、基学习器权重训练。

第一部分:

输入:训练数据集 $T = \{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$;各基学习器损失函数 $\{L(y,f(x))\}$;基学习器集 $\{b(\chi;\gamma)\}$;

输出:各基学习器模型 $\{f(x)\}$。

(1)初始化各 $f(x)$。

(2)针对各个基学习器极小化损失函数:

$$\min(\text{Loss}(y,f(x)))$$

(3)更新基学习器模型参数。

(4)得到 $\{f(x)\}$。

第二部分:

输入:训练数据集合 $T = \{(x_1,y_1),(x_2,y_2),\cdots,$

$(x_N, y_N)\}$;第一部分已经训练完成的基学习器模型,MSE 损失函数;

输出:各基类学习器权重参数。

(1)初始化权重矩阵 W,初始化多模型融合函数:

$$f_{\text{all}}(x) = w_1 \times f_{\text{LR}}(x) + w_2 \times f_{\text{Tree}}(x) + w_3 \times f_{\text{svm}}(x)$$

(2)目标函数:

$$\arg\max_{w} \prod_{i=1}^{N} p(x_i \mid w)$$

(3)最终输出各基学习器参数与对应权重。

3　实　验

3.1　数据分析

文中使用泰坦尼克号之灾数据集验证算法效果。泰坦尼克号之灾是 Kaggle 上经典的二分类问题,造成海难失事的原因之一是乘客和机组人员没有足够的救生艇。尽管幸存下沉有一些运气因素,但有些人比其他人更容易生存,比如女人,孩子和上流社会,通过分析数据,使用机器学习模型,判断乘客能否存活。通过最终结果表明,该数据集可以有效检验各模型性能对比情况。

泰坦尼克号之灾数据集共有训练数据 891 条,有 12 列属性,其中 Cabin 属性由于缺失值占比过多,将属性值转化为有值(yes),无值(no),同时使用众数补偿 Age 中 Null 值。属性信息如表 1 所示,训练数据如图 1 所示。

表 1　属性列表

属性	类型	非空值个数	初步处理方式
PassengerId	Ordinal	891	舍弃
Survived	categorical	891	Label
Pclass	categorical	891	保留
Name	categorical	891	舍弃
Sex	categorical	891	保留
Age	numerical	714	众数补偿
SibSp	numerical	891	保留
Parch	numerical	891	保留
Ticket	Ordinal	891	保留
Fare	numerical	891	保留
Cabin	numerical	204	转化为 Yes、No 保留
Embarked	categorical	889	保留

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
893	3	s, Mrs. James (Ellen N	female	47	1	0	363272	7		S
894	2	yles, Mr. Thomas Franci	male	62	0	0	240276	9.6875		Q
895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.6625		S
896	3	lrs. Alexander (Helga E	female	22	1	1	3101298	12.2875		S
897	3	ensson, Mr. Johan Cervi	male	14	0	0	7538	9.225		S
898	3	Connolly, Miss. Kate	female	30	0	0	330972	7.6292		Q
899	2	dwell, Mr. Albert Fran	male	26	1	1	248738	29		S
900	3	Mrs. Joseph (Sophie Hal	female	18	0	0	2657	7.2292		C
901	3	Davies, Mr. John Samuel	male	21	2	0	A/4 48871	24.15		S
902	3	Ilieff, Mr. Ylio	male		0	0	349220	7.8958		S
903	1	nes, Mr. Charles Cress	male	46	0	0	694	26		S
904	1	. John Pillsbury (Nelle	female	23	1	0	21228	82.2667	B45	S
905	2	Howard, Mr. Benjamin	male	63	1	0	24065	26		S
906	1	rbert Fuller (Carrie Co	female	47	1	0	.E.P. 573	61.175	E31	S
907	2	Mrs. Sebastiano (Argeni	female	24	1	0	/PARIS 21	27.7208		C
908	2	Keane, Mr. Daniel	male	35	0	0	233734	12.35		Q

图 1　训练数据

针对所有保留属性创建与 label 变量的映射图,直观观察变化关系,剔除无明显相关关系的属性,使用保留属性建立特征集合,对离散特征进行因子化,对连续特征进行归一化操作,最终生成特征变量,部分有效属性与 label 对应关系图如图 2 所示。在图中可以明显观察出 Age、Sex 等变量与 label 相关性强,而变量 Name、Ticket 由于是随机化数据从而导致与 label 无明显关系。

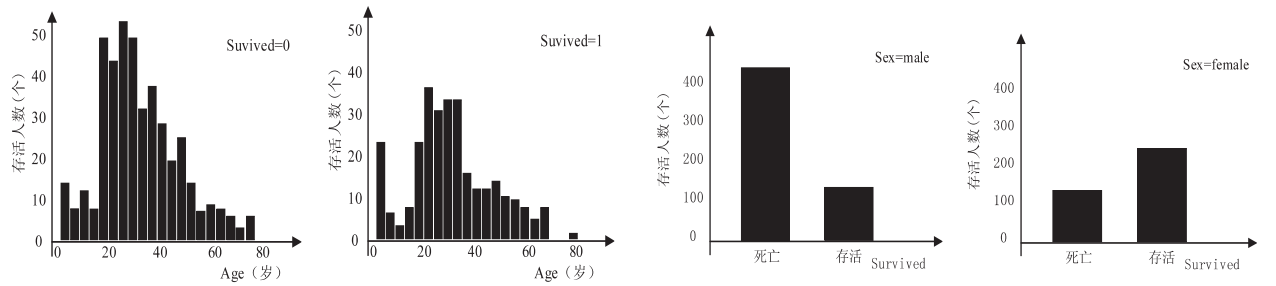

图 2　部分变量与 Label 对应关系

3.2　评价指标

评价模型指标有多类,由于文中为二分类问题,所以选用精确率、召回率、准确率、ROC 评价模型性能。

精确率(Precision)指的是模型输出结果中判断为

正样本的数据中真实为正样本的比例。

召回率(Recall)指的是有多少正样本被准确标出。

设模型输出的正样本集合为 A,真正的正样本集合为 B,则有:

$$\text{Precision}(A,B) = \frac{|A \cap B|}{|A|}$$

$$\text{Recall}(A,B) = \frac{|A \cap B|}{|A|}$$

准确率(Accuracy)衡量的是分类正确的比例。假设是 $\hat{y}$ 是模型输出的预测 label, y 为样本中正确的 label,则准确率为:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y_i} = y_i)$$

ROC 曲线是以假正率为横坐标,真正率为纵坐标的曲线图。设模型预测的正样本集合为 A,真正的正样本集合为 B,所有样本集合为 C,则 A 与 B 的交集个数除以 B 的个数为真正率(true-positive rate), A 与 B 交集的个数除以 C 减 B 的个数为假正率(false-positive rate)。AUC(area under curve)分数是曲线下的面积,越大意味着分类器效果越好。

3.3 实验结果与分析

在表 2 实验数据指标中列举出各个模型在测试集中的评价指标,并增加神经网络与多模型融合进行横向对比,通过对比得出,多模型融合算法在精确率、召回率、准确率、AUC 各个指标上均有明显提升。相对于神经网络这类深度学习模型,多模型融合算法更加适用于小规模数据集。

表 2 实验数据指标

指标	决策树	逻辑回归	SVM	神经网络	多模型融合
Precision	0.805 5	0.750 0	0.666 6	0.722 2	0.805 5
Recall	0.763 1	0.771 4	0.774 1	0.764 7	0.805 5
Accuracy	0.840 0	0.830 0	0.810 0	0.820 0	0.860 0
AUC	0.825 1	0.816 4	0.800 1	0.8065	0.848 0

4 结束语

在小规模数据集合中,多模型融合算法可以融合各个模型优势,对基学习器预测正确结果给予更大权值,对预测错误结果减小权值,通过数据累加,最终增大模型预测准确率,同时提升模型各项指标。相对于深度学习模型需要大量数据进行训练,多模型融合算法更加适用于小数据集。文中在特征选择中并不完善,后续可以通过特征组合等方式进行提升。

参考文献:

[1] 蔡 毅,朱秀芳,孙章丽,等. 半监督集成学习综述[J]. 计算机科学,2017,44(S1):7-13.

[2] 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛,2011,26(3):32-38.

[3] 梁 循. 数据挖掘算法与应用[M]. 北京:北京大学出版社,2006.

[4] RAMAGERI B M. Datamining techniques and applications[J]. Indian Journal of Computer Science & Engineering, 2010,1(4):25-47.

[5] JIANG L, LI C. Scaling up the accuracy of decision-tree classifiers:a Naive-Bayes combination[J]. Journal of Computers,2011,6(7):1325-1331.

[6] HSSINA B,MERBOUHA A,EZZIKOURI H,et al. A comparative study of decision tree ID3 and C4.5[J]. International Journal of Advanced Computer Science & Applications, 2014(2):126-133.

[7] 李孝伟,陈福才,李邵梅. 基于分类规则的 C4.5 决策树改进算法[J]. 计算机工程与设计,2013,34(12):4321-4325.

[8] DOU Huili,WANG Guohua,GUO Min. Algorithm of traffic state probability forecast based on logistic regression[C]// International conference on electronics,information and communication engineering. New York:ASME Press,2012:99-104.

[9] 吴 凯,季新生,刘彩霞. 基于行为预测的微博网络信息传播建模[J]. 计算机应用研究,2013,30(6):1809-1812.

[10] GUPTA A,KUMARAGURU P. Credibility ranking of tweets during high impact events[C]//Proceedings of the 1st workshop in privacy and security in online social media. [s. l.]: ACM,2012:2-12.

[11] 应维云,覃 正,赵 宇,等. SVM 方法及其在客户流失预测中的应用研究[J]. 系统工程理论与实践,2007,27(7):105-110.

[12] 许建华,张学工,李衍达. 支持向量机的新发展[J]. 控制与决策,2004,19(5):481-484.

[13] 王国胜,钟义信. 支持向量机的若干新进展[J]. 电子学报,2001,29(10):1397-1400.

[14] 常继科,赵建辉,任新会,等. 支持向量机综述[J]. 光盘技术,2007(2):4-5.

[15] 李祥纳,艾 青,秦玉平,等. 支持向量机增量学习算法综述[J]. 渤海大学学报:自然科学版,2007,28(2):187-189.