文章编号: 1001-0920(2010)01-0048-05

基于 Bagging 的组合 k-NN预测模型与方法

何 亮, 宋擒豹, 海 振, 沈钧毅
(西安交通大学 电子与信息工程学院, 西安 710049)

摘 要: k-近邻方法基于单一 k 值预测,无法兼顾不同实例可能存在的特征差异,总体预测精度难以保证. 针对该问题,提出了一种基于 Bagging 的组合 k-NN预测模型,并在此基础上实现了具有属性选择的 Bgk-NN 预测方法. 该方法通过训练建立个性化预测模型集合,各模型独立生成未知实例预测值,并以各预测值的中位数作为组合预测结果. Bgk-NN预测可适用于包含离散值属性及连续值属性的各种类型数据集. 标准数据集上的实验表明,Bgk-NN预测精度较之传统 k-NN方法有了明显提高.

关键词: 近邻预测; Bagging; 组合方法
中图分类号: TP274 **文献标识码**: A

Bagging-based ensemble model and algorithm of k-NN prediction

HE Liang, SONG Qin-bao, HAI Zhen, SHEN Jun-yi
(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China. Correspondent: SONG Qin-bao, E-mail: qbsong@mail.xjtu.edu.cn)

Abstract: The existing k-nearest neighbor (k-NN) algorithm predicts in terms of a fixed single k value without considering the diversity of various unknown instances, thus the prediction perpformance can hardly be ensured. Therefore, both an ensemble model of k-NN prediction based on bagging principle and a Bgk-NN prediction algorithm with attributes selection are proposed in this paper. In the novel Bgk-NN algorithm, a set of diverse base k-NN predictors are trained, and the unknown instance is predicted independently by those k-NN predictors. Consequently, the median of all the predicted values is calculated to be the final result of the ensemble model. The Bgk-NN algorithm can fit well all the datasets with no matter discrete or continuous attributes. The experimental results on standard datasets show that, compared with the traditional k-NN predictor, the prediction accuracy of Bgk-NN can be improved effectively.

Key words: k-NN predictor; Bagging; Ensemble method

1 引 言

基于实例学习的 k-NN方法在数据挖掘的分类及数值预测中有着广泛应用,该方法参照训练集中的 k 个近邻实现了未知实例的分类或预测. 关于 k-NN 学习的改进研究,主要集中于提升精度及提高近邻搜索速度方面. 例如动态调整数据集各属性权重[1],基于 Tabu 搜索策略的属性选择及权重分配[2],将不同的距离计算方法加权组合[3],基于本体概念计算样本间的语义距离[4],为高维空间样本建立 B^+ 树索引降低样本搜索维度[5],通过灰关联分析减小近邻搜索空间等方法[6]. 现有的研究虽然在一定程度上改善了 k-NN 方法,但却普遍存在着一个共同问题,即在分类或数值预测过程中对所有未知实例均采用同样数量的近邻作为参照. 然而,该方式很难满足不同实例的个性需求,此情况下仅通过调整距离度量、合理寻找近邻等方法,并不能保证所有实例均获得理想的分类或预测精度.

本文主要基于单一 k 值的 k-NN 数值预测方法的改进进行研究. 不同实例采用不同数量的近邻作为参照的 k-NN 预测方式,能够考虑不同实例的个性化特征,避免传统 k-NN 方法单一 k 值机制的不足. 该预测方式需要预先生成多个不同的 k-NN 预测器. 由于 k-NN 是一种基于实例的学习方法,更适合于数据集较小的应用环境. 如何在小数据集上产

收稿日期: 2009-02-16; **修回日期**: 2009-04-21.
基金项目: 国家自然科学基金重大研究计划项目(90718024); 国家 863 计划项目(2006AA01Z183).
作者简介: 何亮(1975—),男,西安人,博士生,从事数据库及数据挖掘的研究; 宋擒豹(1966—),男,陕西华县人,教授,博士生导师,从事数据挖掘、软件工程等研究.

生具有差异性的若干个不同训练集,并生成不同的个性化 k-NN 预测器,是首先需要解决的问题,组合理论则是实现这一目标的有效途径.

2　组合 k-NN 预测模型

Breiman[7,8]提出的 Bagging 方法是一种独立于特定学习算法的弱学习器提升方法. Bagging 首先在原始数据集上进行反复抽样,生成多个包含重复实例的抽样训练集,并为每个抽样集建立一个基本学习器,将这些弱学习器融合即得到一个组合强学习器. 各抽样集的差异性保证了组合学习器的泛化能力.

为了实现个性化的 k-NN 预测,可通过 Bagging 理论建立多个基本 k-NN 预测模型,这些基本模型的组合能够兼顾不同实例的特征,进行个性化的 k-NN 预测,使总体预测精度的提升成为可能. 基于上述分析,图 1 给出了基于 Bagging 的组合 k-NN 预测模型示意图. 该组合模型由训练数据集抽样、抽样集上的基本 k-NN 预测模型训练和基本 k-NN 模型组合等模块构成.

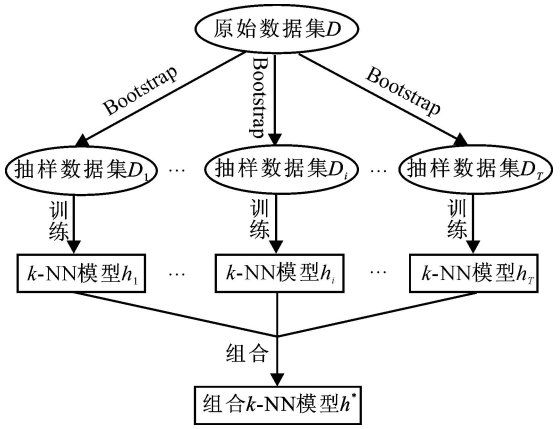

图 1　基于 Bagging 的组合 k-NN 预测模型

组合 k-NN 预测模型首先通过原始数据集 D 上的 Bootstrap 抽样生成各训练子集 D_1, …, D_T,在这些抽样集上进行训练可生成不同的基本 k-NN 预测模型. 其中模型 h_i 应使对应训练集 D_i 中的实例获得相对最优的预测精度,与 D_i 中的实例特征类似的未知实例在 h_i 上获得准确预测的可能性就更大,基本 k-NN 模型集合能够准确预测更多未知实例的可能性也就越大.

基本 k-NN 模型的组合方式决定了组合模型 h^* 对未知实例的最终预测结果,不同组合方式根据各基本 k-NN 模型预测值计算组合预测结果的方法也会不同.

3　Bgk-NN 预测算法

Bgk-NN 预测算法以基于 Bagging 的组合 k-NN 预测模型为基础,建立具有不同偏好的基本 k-NN 预测模型集合,通过组合预测改善传统 k-NN 方法的不足. Bgk-NN 算法可以广泛适用于具有离散值及连续值属性的不同类型数据集.

3.1　算法结构

Bgk-NN 算法由模型训练及未知实例预测两部分构成. 前者主要用于各个基本 k-NN 预测模型的训练,后者给出未知实例的组合预测结果.

图 2 所示为Bgk-NN 算法建立组合 k-NN 预测模型的训练流程,其过程包括数据集抽样和基本模型生成两个阶段. 各训练数据集通过原始数据集上的 Bootstrap 抽样生成,算法训练过程的迭代次数为 T,该值可根据需要预先设定. 抽样前应首先完成数据集的归一化预处理,将各属性映射至相同的值域范围,以保证近邻计算的合理性.

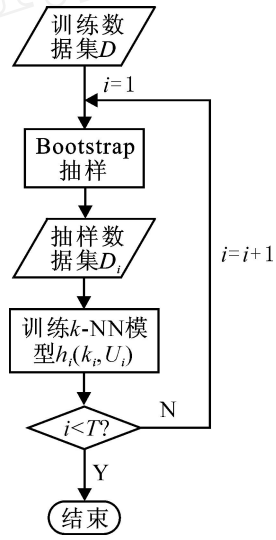

图 2　Bgk-NN 预测算法训练流程

各抽样集上建立的基本 k-NN 预测模型是 Bgk-NN 算法的关键,每个基本 k-NN 模型 h_i ($i \in 1, \cdots, T$) 由参照近邻数 k_i 及属性子集 U_i 两个参数描述,不同抽样集上的模型参数不尽相同,各模型对未知实例的预测结果也会不同. 在 h_i 中使用属性子集 U_i 作为基本 k-NN 模型参数的原因在于,恰当的属性子集可以更合理地度量实例间距离,并可增加 k-NN 算法的不稳定性,使其更适宜于组合理论的应用. 基本模型 h_i 的参数 k_i 及 U_i 应使训练集 D_i 中的实例获得最小的总体预测误差,Bgk-NN 算法采用相对误差作为基本模型的训练标准. Bgk-NN 算法的另一特点在于相互独立的抽样集生成使训练过程可并行进行,显著降低算法的运行时间,提高运行效率.

Bgk-NN 算法的未知实例预测部分主要用于计算各基本 k-NN 模型的独立预测值,并给出组合预测结果. T 个基本预测模型对新实例的预测同样可以并行进行,最终的预测结果经过这些预测值上的

组合函数得出. 以下重点就Bgk-NN 算法中的基本模型训练及组合方式进行分析.

3.2 基本k-NN模型训练

抽样集上的基本k-NN 模型训练是为了选出与训练集相匹配的模型参数,即近邻数 k_i 及属性子集 U_i,两个参数的选择在训练过程中同时进行. Bgk-NN 算法采用基于包装方式的后向消去属性选择方法,该方法依次删除属性集中的一个属性. 重复此过程,直到预测误差不再降低为止,即得到所选的属性子集.

Bgk-NN 算法通过循环方式选出与每个可选 k 值相匹配的属性子集,该属性子集在当前 k 值下可以实现训练集上的最小平均预测误差,k 的可选范围从 1 至预设的最大值 $K_{\max}$. $K_{\max}$ 个平均预测误差中最小值对应的 k 及属性子集,即为抽样集 D_i 上的基本k-NN 预测模型 h_i 的描述参数.

如何确定近邻搜索空间是基本k-NN 模型训练时的另一个重要问题. Bagging 方法通常会直接在各抽样集上进行训练,但对于k-NN 这样的消极学习方法而言,该训练方式却不尽合理. 当采用留一法在抽样集上寻找近邻时,抽样集中存在的重复实例必然可在 $k = 1$ 时实现准确预测. 该现象使得最终选择参数 k 为 1 的可能性很大,但是这样选出的参数缺乏泛化能力,并无实用意义. 因此,基本k-NN 模型的训练应结合原始数据集和抽样数据集进行近邻搜索. Bgk-NN 算法以原始数据集 D 为近邻参照空间逐一搜索抽样集中各实例的近邻,由于抽样集是通过 D 上的Bootstrap 抽样产生的,抽样集中各实例必然也在 D 中存在. 为了避免类似上述问题的出现,Bgk-NN 算法搜索近邻时会先将该实例从 D 中去除.

Bgk-NN 算法基本预测模型训练函数 BaseModel(D_i, D) 的具体实现如下:

1) for $k = 1, \cdots, K_{\max}$;
2) 参照 k 个近邻依次预测 D_i 中各实例,计算平均相对误差 me;
3) for $u = 1, \cdots, |U|$, //$|U|$ 为 U 的属性数,初始条件下 U 为 D 的属性全集;
4) $U' = U - U(u)$; //去除 U 的第 u 个属性;
5) $D_i' = D_i(U')$; //仅保留 D_i 的 U' 属性;
6) $D' = D(U')$; //仅保留 D 的 U' 属性;
7) 以 D' 中的 k 个近邻为参照预测 D_i' 中各实例,记平均预测误差为 $\varepsilon(u)$;
8) end for
9) $\varepsilon = \min(\varepsilon(u))$, $u = 1, \cdots, |U|$;
10) if $\varepsilon \leq me$ then
11) $U = U_{\min}$; //$U_{\min}$ 为 ε对应的属性集;
12) $me = \varepsilon$,转至 3); //继续属性子集选择;
13) else
14) me 即为 k 对应的最小预测误差,关于 k 的属性子集选择结束;
15) end if
16) $e(k) = me$;
17) end for
18) $k_i = \arg\min_k(e(k))$, $k = 1, \cdots, K_{\max}$;
19) 返回模型 h_i 的近邻值 k_i,属性子集 U_i.

函数返回面向 D_i 的近邻数 k_i 和属性子集 U_i. 步骤 1) ～ 17) 依次计算每个备选 $k(k \in 1, \cdots, K_{\max})$ 值上的最优属性子集及平均误差;步骤 18) 从 $K_{\max}$ 个平均误差中选出最小误差对应的 k 值及属性子集作为函数返回结果. 步骤 2) 首先计算出未进行属性选择前以 k 为近邻时的预测误差 me,其目的在于更稳妥的模型训练. 经过初次属性选择,如果在属性子集上的预测误差大于 me,则意味着以 k 为参数时在属性子集上搜索近邻并未提升k-NN 的预测精度. 此时后向消去过程即可结束,并将属性全集作为模型参数. 步骤 3) ～ 8) 为生成属性子集的过程,逐一删除属性集 U 中的每个属性并计算平均误差,起始情况下每个 k 值对应的属性集 U 均为 D 的属性全集. 随着属性选择的进行,U 逐渐减小直至误差无法继续降低为止. 产生最小误差的属性子集(属性数为 $|U| - 1$) 通过步骤 9) 选出,并与上一次选择过程产生的最小误差 me(此时属性数为 $|U|$) 比较,me 最初为无属性选择的 k 近邻预测误差,随着循环进行,me 被赋值为上一属性选择过程的最小误差. 如果新一轮属性选择后误差进一步减小,则继续属性子集选择,否则选择过程结束,该功能由步骤 10) ～ 15) 实现. 每个 k 值的属性子集选择过程结束后,该 k 值在训练集 D_i 上的最小误差同时产生,数组 $e(\ )$ 专用于记录不同 k 值的最小误差.

3.3 模型的组合

训练产生的各基本模型 $h_1, \cdots, h_T$ 分别生成未知实例 (x_s, y_s) 的预测值 $\hat{y}_1, \cdots, \hat{y}_T$. Bg$k$-NN 以各基本预测值的中位数作为组合预测结果,中位数方式可综合各个基本模型的预测偏好且便于计算. 也可通过加权中位数组合方式强化训练精度较理想的基本k-NN 模型的预测能力,各模型权重由平均训练

误差决定. 如果某个未知实例与高权重模型对应训练集中的实例特征类似,则组合模型能够准确预测该实例的可能性同样也较大,但是对于特征差异较大的实例其预测准确性则会受到影响. 简单的算术平均也是一种可选的处理方式,不过该方式易受基本模型预测值中噪声的影响. 综合以上分析, Bgk-NN 算法采用中位数组合方式.

3.4　算法实现

Bgk-NN 算法的各重要步骤均已详细分析,最后给出完整的Bgk-NN 预测算法如下:

输入:训练数据集 D,迭代次数 T,待预测实例 x_s;

输出:实例 x_s 的预测值 $\hat{y}_s$.

1) 初始化实例权重

$w(i) = 1/n,\ i = 1, \cdots, n,\ n = |D|$;

2) for $i = 1, \cdots, T$;

3) 根据 w 对原始数据集 D 进行 Bootstrap 抽样,得到训练集 D_i, $|D_i| = n$;

4) 调用函数 BaseModel(D_i, D) 建立 D_i 上的基本 k-NN 预测模型 $h_i: (k_i, A_i)$;

5) end for　// 训练过程结束;

6) for $i = 1, \cdots, T$;

7) 采用模型 h_i 预测 x_s,生成预测值 $\hat{y}_i$;

8) end for　// 生成各基本模型预测值;

9) $\hat{y}_s = \text{median}(\hat{y}_1, \cdots, \hat{y}_T)$　// 生成组合预测结果.

算法由 4 部分构成,首先初始化训练集中各实例权重,赋予每个实例同样的权重 $1/n$;接下来需要完成数据集上的抽样及基本模型训练,由步骤 2) ～ 5) 实现,等概率抽样使得训练集 $D_1, \cdots, D_T$ 不会存在倾向性,可考虑不同类型的实例分布;算法从步骤 6) 开始未知实例的预测,每个基本模型 h_i 生成一个预测值 $\hat{y}_i$;最后算法在步骤 9) 返回组合预测结果 $\hat{y}_s$,其中 median( ) 为中位数函数.

4　实验结果

本文将Bgk-NN 预测方法与具有属性选择的单一 k 值k-NN 预测方法进行实验对比. 实验所用 auto93, auto Horse, pollution 标准数据集[10] 由不同的连续值及离散值属性构成,涉及汽车及环境保护领域. 各数据集均进行了归一化及去除信息缺失实例等预处理操作. 实验采用 3 折交叉验证方法进行,为避免分折及 Bootstrap 抽样中的偶然因素影响,实验在各数据集上重复 3 次,分别就两种预测方法的最小误差、最大误差、平均误差进行比较. 算法基于 Java 实现, 硬件环境:CPU 为 Intel Core2 Duo E4500, 2.20 GHz,内存 1 GB,操作系统 Windows XP.

表 1 为 auto93 数据集误差统计,该数据集包含 82 个实例、23 个属性. 实验中传统k-NN 方法的各种误差获得了不同程度的降低. 以 $T = 20$ 为例, Bgk-NN 的平均误差较传统 k-NN 预测降低了 20.75 %～22.52 %. 实验所用Bgk-NN 算法采用顺序执行方式实现,因此随着迭代次数 T的增大,时间有增加趋势.

表 1　auto93 数据集预测误差统计

次数	实验内容	k-NN	Bgk-NN			
			$T = 5$	$T = 10$	$T = 15$	$T = 20$
1	Min/ %	0.33	0.16	0.11	0.03	0.12
	Max/ %	92.84	68.79	68.70	68.22	63.50
	Avg/ %	22.07	17.93	17.91	17.67	17.42
	Time/ s	10.97	64.86	130.24	190.08	255.57
2	Min/ %	1.08	0.90	0.36	0.19	0.19
	Max/ %	133.00	61.56	61.56	59.09	57.62
	Avg/ %	23.52	19.21	19.15	18.83	18.71
	Time/ s	11.34	38.53	67.02	131.25	192.25
3	Min/ %	0.89	0.41	0.28	0.22	0.16
	Max/ %	123.18	71.04	55.13	52.02	50.48
	Avg/ %	22.29	18.43	18.11	17.65	17.27
	Time/ s	13.17	69.08	141.11	209.45	274.03

表 2 为 auto Horse 数据集实验结果,该数据集含有 159 个实例、26 个属性. 各次实验的最小误差为 0 及最大误差不随 T 值变化的现象,与该数据集存在部分实例在待预测属性上取值相同的情况有关. $T = 20$ 时,Bgk-NN 的平均误差相对基本k-NN 预测降低了 19.49 %～23.72 %. 由于该数据集属性及实例数相对较多,因此其运行时间较长.

表3给出了pollution数据集的预测误差统计,

表 2　auto Horse 数据集预测误差统计

次数	实验内容	k-NN	Bgk-NN			
			$T = 5$	$T = 10$	$T = 15$	$T = 20$
1	Min/ %	0	0	0	0	0
	Max/ %	58.82	45.83	45.83	44.97	44.38
	Avg/ %	4.32	3.88	3.79	3.39	3.41
	Time/ s	33.63	175.08	327.78	490.11	627.51
2	Min/ %	0	0	0	0	0
	Max/ %	55.88	53.19	43.75	43.75	41.67
	Avg/ %	6.45	5.13	5.00	4.91	4.92
	Time/ s	26.55	142.77	292.80	458.61	619.55
3	Min/ %	0	0	0	0	0
	Max/ %	45.83	45.83	45.83	45.83	45.50
	Avg/ %	5.90	5.20	5.17	5.14	4.75
	Time/ s	29.42	152.41	313.47	458.19	588.99

该数据集含有 60 个实例、16 个属性. Bgk-NN 方法在该数据集上的各类误差相对基本k-NN 方法有不同程度的降低. T = 20 时,Bgk-NN 方法的平均误差较基本k-NN 预测降低了 12.08 %～21.93 %.

表 3 pollution 数据集预测误差统计

次数	实验内容	k-NN	Bgk-NN			
			T = 5	T = 10	T = 15	T = 20
1	Min/ %	0.14	0.12	0.04	0.14	0.11
	Max/ %	17.44	15.19	13.87	13.64	13.27
	Avg/ %	4.97	4.13	3.96	3.89	3.88
	Time/ s	0.59	14.53	28.91	42.97	57.25
2	Min/ %	0.21	0.21	0.18	0.12	0.03
	Max/ %	12.39	11.96	10.37	10.66	10.37
	Avg/ %	4.47	4.18	4.04	3.98	3.93
	Time/ s	0.58	14.73	29.09	43.95	58.52
3	Min/ %	0.33	0.18	0.24	0.01	0.01
	Max/ %	15.88	12.98	12.48	11.76	11.67
	Avg/ %	4.29	3.88	3.76	3.71	3.70
	Time/ s	0.59	14.70	29.10	43.34	58.16

Bgk-NN 算法的预测误差随 T 值增大呈现的减小趋势与组合模型的特点一致,T 越大相当于为不同实例建立的个性化模型更具有针对性,预测精度自然更为理想. 不同基本k-NN 模型间的差异程度决定了Bgk-NN 算法与传统k-NN 算法间的预测精度差异. 各抽样集的相似度越低,训练得到的各基本模型的个性化特征也就越显著,可以适应的未知实例也就越多. 此时,Bgk-NN 的总体预测精度较之传统k-NN 方法更为理想. 如果基本模型的数量较少或差异性不够显著,则Bgk-NN 改善k-NN 预测精度的显著性也会随之降低.

对于顺序执行的Bgk-NN 算法而言,T 的增加是以时间开销为代价的,可综合组合预测方法的精度和时间要求选择合适的 T 值,也可通过调整属性选择方法缩短Bgk-NN 的时间开销. Bgk-NN 采用的后向消去属性选择过程占据了算法整体开销的重要部分,在尽可能不损失预测精度的前提下,通过优化属性选择方式可进一步提升Bgk-NN 的运行效率,如采用基于过滤(filter) 方式的属性选择. 此外,为实例建立索引加快近邻搜索速度也可在一定程度上降低算法的时间开销.

5 结 论

本文提出的基于 Bagging 的组合k-NN 预测模型及Bgk-NN 算法,为传统k-NN 方法预先建立组合k-NN 预测模型集合,以组合预测值作为未知实例的最终预测结果. 基于组合理论的k-NN 预测能够考虑不同实例的个性化特征,避免了传统k-NN 方法对所有实例均采用单一 k 值进行预测,难以兼顾不同实例特征的不足. Bgk-NN 算法不受数据集属性类型的限制,可以广泛应用于具有连续值属性及离散值属性的各类数值预测问题.

参考文献(References)

[1] Han E H, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classify cation [C]. Proc of the 5th Pacific-Asia Conf on Knowledge Discovery and Data Mining. London: Springer Verlag, 2001: 53-65.

[2] Tahir M A, Bouridane P, Kurugollu F. Simultaneous feature selection and feature weighting using hybrid tabu search/ k-nearest neighbor classifier [J]. Pattern Recognition Letters, 2007, 28(4): 438-446.

[3] Yamada T, Yamashita K, Ishii N, et al. Text classification by combining different distance functions with weights[C]. Proc of the 7th ACIS Int Conf on Software Engineering, Artificial Intelligence, Networking, and Parallel/ Distributed Computing. Los Alamitos: IEEE Computer Society, 2006: 85-90.

[4] 杨立, 左春, 王裕国. 基于语义距离的 k-最近邻分类方法[J]. 软件学报, 2005, 16(12): 2054-2062.
(Yang L, Zuo C, Wang Y G. k-nearest neighbor classification based on semantic distance [J]. J of Software, 2005, 16(12): 2054-2062.)

[5] Jagadish H V, Ooib B C, Tan K L, et al. iDistance: An adaptive B^+-tree based indexing method for nearest neighbor search[J]. ACM Trans on Database Systems, 2005, 30(2): 364-397.

[6] Huang P C. A novel gray-based reduced NN classification method[J]. Pattern Recognition, 2006, 39(11): 1979-1986.

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[8] Breiman L. Bias, variance, and arcing classifiers[R]. Berkeley: University of California, 1996.

[9] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. J of Machine Learning Research, 2003, 3: 1157-1182.

[10] Department of Computer Science, University of WAIKATO. Weka Machine Learning Project [DB/ OL]. (2008-05-10). http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html.