第43卷 第4期 西 安 交 通 大 学 学 报 Vol. 43 №4
2009年4月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Apr. 2009

一种新的组合 k-近邻预测方法

何亮，宋擒豹，沈钧毅，海振
（西安交通大学计算机科学与技术系，710049，西安）

摘要：针对传统 k-近邻（k-NN）算法基于单一 k 值预测难以兼顾不同样本的个性，从而导致总体预测精度不够理想的问题，提出了一种组合 Bk-NN 预测方法. 首先通过 Boosting 理论建立了个性化预测模型集，然后分别采用每个模型对样本进行独立预测，最后各模型预测值的加权和将作为最终预测结果. Bk-NN 预测充分考虑了不同类型的样本可能要求不同的预测模型与之相适应的情况，有效降低了预测误差. 与其他方法不同的是，Bk-NN 预测对数据集的属性类型没有特殊要求. 在标准数据集上的实验结果表明，Bk-NN 预测精度比传统 k-NN 方法平均提高了 6.44%～15.25%.

关键词：近邻算法；预测模型；Boosting 理论；组合方法
中图分类号：TP274 **文献标志码**：A **文章编号**：0253-987X(2009)04-0005-05

New Combined k-Nearest Neighbor Predictor

HE Liang, SONG Qinbao, SHEN Junyi, HAI Zhen
(Department of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: The prediction accuracy of various instances can hardly be ensured by the existing k-nearest neighbor (k-NN) predictor, which works under a single k value. A combined k-NN algorithm, Bk-NN predictor, is proposed in this paper. The novel Bk-NN algorithm is used to build up a set of prediction models based on the Boosting principle, and then each model is used to predict a new instance respectively. The final predicted value of this instance is the weighted sum of these prediction values. The prediction error is reduced since the circumstance that various instances demand specific prediction models to match with is thoroughly taken into account by the Bk-NN algorithm. Moreover, the Bk-NN predictor can work well with both discrete and continuous attribute values. The experimental results on standard datasets show that, compared with the traditional k-NN predictor, the prediction accuracies of the Bk-NN predictor are improved by 6.44%-15.25%.

Keywords: nearest neighbor algorithm; prediction model; Boosting principle; combined method

基于实例学习的 k-邻近（k-NN）算法是一种被广泛使用的数据挖掘分类及预测方法. 关于 k-NN 方法的改进研究主要集中于，通过调整距离度量方式来提升分类准确率及加快近邻搜索速度等方面[1-4]. 这些改进的共同点在于，任何待分类样本参照的近邻数，即 k 值是固定不变的，并未考虑不同样本在同一 k 值下的分类准确性的差异，这种单一 k 值机制很难满足不同样本的个性需求，因此也难以保证各样本均获得理想的分类精度. 对于 k-NN 数值预测，由于预测属性是连续型数值，较之分类中的类标号更加多样化，因此同一样本的预测值往往随 k 的变化而变，预测结果受 k 值的影响更加突出.

本文将 k-NN 预测与 Boosting 理论相结合，提出了基于组合模型的 Bk-NN（Boosted k-NN）预测方法.

收稿日期：2008-09-23. 作者简介：何亮（1975—），男，博士生；宋擒豹（联系人），男，教授，博士生导师. 基金项目：国家自然科学基金重大研究规划资助项目（90718024）；国家高技术研究发展计划资助项目（2006AA01Z183）.

1 Boosting

Boosting 理论是通过多个基本学习器的组合提升弱学习器的性能. 该方法在原始训练数据集上重复带权重 Bootstrap 抽样过程生成多个新训练数据集,每个数据集将会产生一个基本学习器,在已生成的学习器上预测误差越大的样本将获得越高的权重,因此新抽样的数据集总是集中于那些难度较高的样本,而该数据集上的基本学习器也是针对这些样本进行训练的. Boosting 方法经过迭代地抽样、训练、调整样本权重过程,可建立组合模型,组合模型的性能将优于任何一个基本学习器.

Boosting 方法中最具代表性的是 Freund 和 Schapire 提出的二元分类自适应 AdaBoost 算法及多元分类的 AdaBoost. M1 和 AdaBoost. M2 算法[5]. 此后,AdaBoost 算法簇中还陆续出现了 AdaBoost. OC 及 AdaBoost. MO 等基于纠错输出编码的多元分类的 AdaBoost 算法. Boosting 方法不仅可以提高弱分类器性能[6-7],而且能提升数值预测算法的性能. 另有学者也提出了多种用于数值预测问题的 Boosting 算法[5, 8-10].

2 基于 Boosting 的组合 k-NN 预测方法

2.1 组合 k-NN 预测模型

组合 k-NN 预测模型由若干个基本 k-NN 模型构成,其建立过程主要包括如下 2 个步骤.

(1)在原始训练数据集上采用 Bootstrap 抽样产生新的训练子集,抽样概率与样本权重分布保持一致. 初始情况下,原始数据集中各样本权重均为 $1/N$(N 为样本总数),随着新预测模型的建立,各样本的权重也将随之改变.

(2)建立适合当前抽样数据集的 k-NN 预测模型. 建立模型的过程就是选择 k 值的过程,由于抽样数据集互不相同,每个预测模型的 k 值也会不同.

新的基本预测模型建立之后,原始训练数据集中各样本的权重也要随之调整.

图 1 给出了建立组合 k-NN 预测模型的过程,其中 D 为原始训练数据集,t 为迭代次数(初值为 1),$\boldsymbol{w}_t$ 为第 t 次迭代时 D 的样本权重向量. 根据 $\boldsymbol{w}_t$ 抽样得到训练集 D_t 并建立预测模型 h_t,再根据 h_t 的预测精度调整 $\boldsymbol{w}_t$,得到新的权重向量 $\boldsymbol{w}_{t+1}$,如此循环,直至满足迭代终止条件,因此得到一组 k-NN 预测模型 $h_1,\cdots,h_T$.

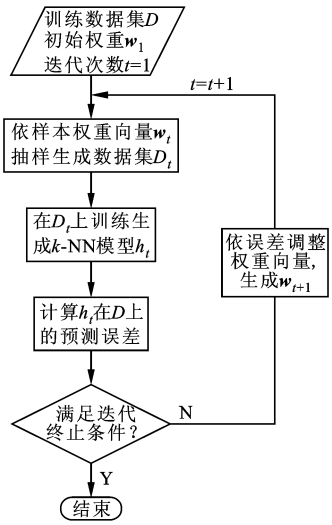

图 1 组合 k-NN 预测模型建立流程图

各基本模型 $h_t(t=1,\cdots,T)$ 对新样本 $(\boldsymbol{x}_p, y_p)$ 均有其预测值,通过模型系数 $C_t(t=1,\cdots,T)$ 将 T 个预测值组合即可获得最终预测结果.

2.2 h_t 的建立

模型 h_t 主要由 k 值决定,h_t 对训练集 D_t 中的样本$(\boldsymbol{x},y)$进行预测时,以该样本到 k 个近邻$(\boldsymbol{x}_1, y_1),\cdots,(\boldsymbol{x}_k, y_k)$的距离为权重,通过计算 $y_1,\cdots,y_k$ 的距离加权和得到$(\boldsymbol{x},y)$的预测值,距离$(\boldsymbol{x},y)$越近的样本权重越高. k 的变化将直接影响 D_t 中各样本的预测精度,为模型 h_t 选择的 k 值应使该模型在训练数据集 D_t 上获得相对最优的平均预测精度.

模型 h_t 在预测 D_t 中的样本$(\boldsymbol{x},y)$时需要参照近邻样本,如何寻找$(\boldsymbol{x},y)$的近邻也是一个非常重要的问题. 原则上$(\boldsymbol{x},y)$参照的 k 个近邻应在原始训练集 D 中寻找,但是由于 D_t 是在 D 上抽样产生的,所以当$(\boldsymbol{x},y)\in D_t$ 时必定也有$(\boldsymbol{x},y)\in D$,即此时$(\boldsymbol{x},y)$必然会同时出现在自己的近邻样本集合中,这将干扰 h_t 选择合理的 k 值. 因此,在 D 中寻找$(\boldsymbol{x},y)$的近邻时应首先将$(\boldsymbol{x},y)$自身从 D 中去除.

设 D 的样本数为 N,由 Bootstrap 抽样方法可知所有抽样集的样本数同样为 N. 基本 k-NN 预测模型 h_t 由函数 FindK(D_t, D)建立,具体过程如图 2 所示.

函数返回的向量 $\boldsymbol{E}_t$ 表示模型 h_t 对原始训练数据集 D 中各样本采用留一法预测时的相对误差,调整样本权重分布时将会使用 $\boldsymbol{E}_t$.

2.3 样本权重的调整

在基于 Boosting 的组合预测方法中,损失函数是权重调整时经常使用的度量依据. 样本的损失越大,则表示预测误差越大,此时应增加样本权重,以

Function FindK(D_t, D)

1 for k=1$\cdots K_{\max}$ //$K_{\max}$为预置最大近邻数

2 参照k个近邻预测D_t中各样本,记第i个样本的相对误差为$\varepsilon(i)$, i=1,$\cdots$,N;

3 计算近邻数为k时的平均预测误差

$$e(k)=(\sum_{i=1}^{n}\varepsilon(i))/N$$

4 end for

5 $K_t=\underset{k}{\operatorname{argmin}}(e(k)),k=1,\cdots,K_{\max}$;

6 返回模型h_t的k值K_t、h_t在D上的预测误差向量$\boldsymbol{E}_t$;

图 2　函数 FindK(D_t, D)的实现过程

便再次抽样时更容易选中该样本；样本的损失越小，则意味着预测误差越小，此时可以降低样本的权重.

经典数值预测算法 AdaBoost. R[5]是以均方误差作为损失度量标准的，该算法将预测问题映射为无限多个二元分类问题的集合，每个二元分类问题表示待预测属性值域中的某个值 y 与当前样本预测值 $h(x)$之间的大小关系，当 $y\geqslant h(x)$时该分类器返回 1，否则返回 0. 由于值域是连续区间，因此映射得到的分类问题将是无限多个. Freund 和 Schapire 已经证明，最小化这些分类器的分类误差与最小化原始问题中预测器的均方误差损失在效果上是等价的，因此 AdaBoost. R 算法可以通过最小化分类误差的方式提升预测算法的精度. 选择不同的损失函数，分类器损失的计算随之不同. AdaBoost. R 算法以分类器中的映射样本为权重调整对象，需要进行大量的分段积分运算，不易于实现. 为了简化算法复杂度，文献[11]在 AdaBoost. R 基础上进行了改进，并引入了不同类型的损失函数.

本文在调整各样本权重的过程中参照了文献[11]的方法，具体地，样本$(\boldsymbol{x},y)$在模型 h_t 上的损失函数采用了以下计算式

$$L_1=\frac{|\hat{y}-y|}{S} \tag{1}$$

$$L_2=\frac{|\hat{y}-y|^2}{S^2} \tag{2}$$

$$L_3=1-\exp\left(\frac{-|\hat{y}-y|}{S}\right) \tag{3}$$

式中：$S=\sup|\hat{y}_i-y_i|$, $i=1,\cdots,N$. 显然，$L_1,L_2,L_3\in[0,1]$. 损失函数首先将样本预测误差映射到[0,1]区间，然后按照以下加权平均损失计算式调整各样本权重

$$\bar{L}_t=\sum_{i=1}^{N}\boldsymbol{w}_t(i)L_t(i) \tag{4}$$

$$\beta_t=\frac{\bar{L}_t}{1-\bar{L}_t} \tag{5}$$

$$\boldsymbol{w}'_{t+1}(i)=\boldsymbol{w}_t(i)\beta_t^{1-L_t(i)} \tag{6}$$

$$\boldsymbol{w}_{t+1}(i)=\boldsymbol{w}'_{t+1}(i)/\sum_{i=1}^{N}\boldsymbol{w}'_{t+1}(i) \tag{7}$$

式(4)～式(7)中：$\boldsymbol{w}_t(i)$、$L_t(i)$分别表示数据集 D 的第 i 个样本在模型 h_t 上的权重及损失；$\bar{L}_t$ 为 D 的加权平均损失；β_t 表示模型 h_t 的预测置信度，加权损失越小，则 β_t 越小，预测的置信度越高.

利用式(6)可计算调整后的权重分布，利用式(7)可对新的权重进行归一化处理，训练子集 D_{t+1} 是按照 $\boldsymbol{w}_{t+1}$ 在 D 上抽样产生的. 当加权平均损失 $\bar{L}_t>0.5$ 时，相当于映射分类器的分类精度已经劣于随机猜测，此时可终止迭代过程.

2.4　基本模型的组合

依据组合 k-NN 模型中的各基本模型 $h_1,\cdots,h_T$ 分别生成测试样本$(\boldsymbol{x}_s,y_s)$的预测值 $\hat{y}_1,\cdots,\hat{y}_T$，这些预测值的加权和就是组合模型的最终预测结果 $\hat{y}_s$，$\hat{y}_t$ 的权重为 h_t 的系数 C_t. 模型系数在很大程度上影响组合模型的预测精度，计算各模型系数时应充分考虑不同模型对不同样本的预测精度的差异.

为了确定 C_t，首先需要在原始训练数据集 D 中寻找$(\boldsymbol{x}_s,y_s)$的最近邻样本$(\boldsymbol{x}_{\text{near}},y_{\text{near}})$，$C_t$ 的大小将由$(\boldsymbol{x}_{\text{near}},y_{\text{near}})$在 h_t 上的误差决定. $(\boldsymbol{x}_{\text{near}},y_{\text{near}})$在 h_t 上的预测误差越小，h_t 准确预测$(\boldsymbol{x}_s,y_s)$的可能性越大，此时应给予 h_t 较高的权重系数；反之，C_t 应相对较低，而归一化后的各预测误差倒数即为各模型系数 C_t. 不同测试样本的最近邻会不相同，各模型的系数也随之改变，这样对于不同的测试样本，均有可能建立合理的模型组合比例.

Bk-NN 预测方法如图 3 所示.

对于含有 N 个样本的训练数据集，传统 k-NN 算法通过留一法寻找合适的 k 值，其计算复杂度为 $O(N^2)$，而 Bk-NN 方法需要建立 T 个基本预测模型，故其复杂度为 $O(TN^2)$，但样本权重调整等开销相对于基本模型生成可以忽略.

3　实验与分析

3.1　实验设置

本文将 Bk-NN 与传统 k-NN 预测方法进行了实验对比分析. 实验所用 auto93、autoHorse、autoMpg 数据集涉及汽车相关领域，来源于开源数据挖掘工具 WEKA 网站提供的压缩文件包[12]. 各数据集均进行了归一化及去除信息缺失样本等预处理

算法：基于组合理论的Bk-NN预测算法
输入：训练数据集D,测试样本$(\boldsymbol{x}_s, y_s)$
输出：样本$(\boldsymbol{x}_s, y_s)$的预测值$\hat{y}_s$

```
t=1;
初始化样本权重
    w_t(i)=1/N, i=1,…,N, N=|D|;
L̄_t=0;
while L̄_t≤0.5 {
    根据w_t对数据集D进行Bootstrap抽样,得到数据集D_t,|D_t|=N;
    针对D_t建立预测模型h_t
    (K_t, E_t)=FindK(D_t, D);
    按式(1)计算数据集D中各样本损失L_t(i),i=1,…,N;
    按式(4)计算数据集D的加权平均损失L̄_t;
    if L̄_t≤0.5 then
        按式(5)～式(7)调整D中各样本权重并归一化;
        t=t+1;
    end if }
T=t-1;      //T为基本预测模型数
for t=1,…,T
    计算样本(x_s, y_s)在模型h_t上的预测值y_pre(t)
end for
计算各模型系数C_t, t=1,…,T;
ŷ_s=Σ_{t=1}^{T} C(t)y_pre(t)
```

图3 Bk-NN预测方法

操作.预处理后的auto93数据集包含82个样本,23个属性;autoHorse数据集包含159个样本,26个属性;autoMpg数据集包含390个样本,8个属性.实验采用3折交叉验证方法进行,由于交叉验证数据分折及Boosting方法样本重抽样均具有随机性,为了检验数据集不稳定性对预测方法的影响并避免一次实验中偶然因素的干扰,本实验在各训练数据集上重复进行3次,分别就2种预测方法的最小误差、最大误差及平均误差进行对比.算法基于Java语言实现,实验硬件环境:CPU为Intel Core2 Duo E4500,2.20 GHz,内存1 GB,操作系统Windows XP.

3.2 结果及分析

表1记录了auto93数据集的预测误差统计,所有误差均为相对误差.可以看到,Bk-NN的最小误差、最大误差及平均误差较传统k-NN均有不同幅度的降低.3次实验中k-NN的预测最小误差在0.70%～1.08%之间,虽然该误差比较理想,但是Bk-NN的预测误差较之仍然显著降低,其最小误差不超过0.54%,同时Bk-NN的最大误差同样如此,其中最后一次实验效果更为显著.总体上,Bk-NN的平均误差相对k-NN降低了13.77%～15.25%.由于2种方法的运行时间均不超过0.3 s,因此Bk-NN算法增加的计算开销实际上是难以分辨的.

表1 auto93数据集预测误差统计

实验次数	方法	最小误差/%	最大误差/%	平均误差/%	运行时间/s
1	k-NN	1.08	97.03	24.27	0.032
	Bk-NN	0.54	71.21	20.57	0.281
2	k-NN	0.70	84.57	22.75	0.047
	Bk-NN	0.26	69.57	19.36	0.297
3	k-NN	0.82	93.56	20.19	0.025
	Bk-NN	0.31	51.02	17.41	0.188

表2所示为autoHorse数据集实验结果.由于该数据集有部分样本在预测属性上取值相同,因此其总体预测精度较好,2种方法均实现了一些样本的无误差预测.虽然最大误差的降幅不如auto93数据集那样显著,但是平均预测精度仍然获得了9.52%～13.59%的提升.

表2 autoHorse数据集预测误差统计

实验次数	方法	最小误差/%	最大误差/%	平均误差/%	运行时间/s
1	k-NN	0	45.83	8.40	0.078
	Bk-NN	0	42.38	7.51	0.812
2	k-NN	0	47.16	7.43	0.111
	Bk-NN	0	43.79	6.42	1.156
3	k-NN	0	42.17	7.67	0.109
	Bk-NN	0	40.66	6.94	1.203

表3为autoMpg数据集的误差统计,传统k-NN在各次实验中存在最小误差,而Bk-NN充分发挥了组合方法的特点,其将部分样本的预测误差降至为0,同时最大误差也有降低趋势.Bk-NN在该数据集上的平均性能提升不如前2个数据集显著,其在6.44%～6.72%间浮动,但2种方法的绝对运行时间仍然很短.

表3 autoMpg数据集预测误差统计

实验次数	方法	最小误差/%	最大误差/%	平均误差/%	运行时间/s
1	k-NN	0.08	59.85	11.02	1.032
	Bk-NN	0.00	52.63	10.31	18.656
2	k-NN	0.02	58.23	10.72	1.016
	Bk-NN	0.00	40.62	10.01	14.218
3	k-NN	0.11	73.09	10.57	1.047
	Bk-NN	0.00	48.78	9.86	11.781

4 结 论

本文提出的基于 Boosting 的组合 k-NN 预测方法可针对训练数据集中不同样本建立个性化预测模型，新样本的预测以不同个性化模型预测值的加权和作为最终预测值，与当前待预测样本特性相适应的个性化模型将获得较高权重，从而保证了该样本的预测精度. Bk-NN 方法采用组合模型进行预测，避免了传统 k-NN 方法难以保证所有样本均获得较高预测精度的局限性. 由于 Bk-NN 方法不受数据集属性类型的限制，因此它与仅能处理连续数值属性数据的回归预测方法相比有着更广泛的应用领域.

参考文献：

[1] HAN E H, KARYPIS G, KUMAR V. Text categorization using weight adjusted k-nearest neighbor classification[C] // Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany: Springer-Verlag, 2001:53-65.

[2] YAMADA T, YAMASHITA K, ISHII N, et al. Text classification by combining different distance functions with weights[C] // Proceedings of the 7th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Los Alamitos, CA, USA: IEEE Computer Society, 2006: 85-90.

[3] 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 Li, ZUO Chun, WANG Yuguo. K-nearest neighbor classification based on semantic distance [J]. Journal of Software, 2005,16(12):2054-2062.

[4] JAGADISH H V, OOI B C, TAN K L, et al. iDistance: an adaptive B^+-tree based indexing method for nearest neighbor search[J]. ACM Transactions on Database Systems, 2005, 30(2):364-397.

[5] FREUND Y, SCHAPIRE R E, A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1):119-139.

[6] FERN X Z, BRODLEY C E, Boosting lazy decision trees [C] // 20th International Conference on Machine Learning. Menlo Park, CA, USA: American Association for Artificial Intelligence, 2003: 178-185.

[7] 张雷，李人厚. 基于免疫原理和 Boosting 机制的模糊分类规则挖掘算法[J]. 西安交通大学学报，2007，41(8)：927-930.
ZHANG Lei, LI Renhou. Algorithm for mining fuzzy classification rules based on immune principles and boosting mechanism [J]. Journal of Xi′an Jiaotong University, 2007,41(8):927-930.

[8] BREIMAN L. Prediction games and arcing algorithms [J]. Neural Computation, 1999, 11(7):1493-1517.

[9] RIDGEWAY G, MADIGAN D, RICHARDSON T. Boosting methodology for regression problems[C] // Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999: 152-161.

[10] KÉGL B. Robust regression by boosting the median [C]//16th Annual Conference on Learning Theory and 7th Kernel Workshop. Berlin, Germany: Springer-Verlag, 2003: 258-272.

[11] DRUCKER H. Improving regressors using Boosting techniques[C] // The 14th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997:107-115.

[12] University of Waikato. Weka machine learning project [EB/OL]. [2008-05-10]. http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html.

（编辑　苗凌）