文章编号：2095-2163(2021)03-0138-05 中图分类号：TP399 文献标志码：A

基于改进 KD 树的 k 近邻算法在欺诈检测中的应用

吴金娥，段倩倩
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要： 面对互联网交易中店家靠刷销量欺骗消费者的问题，提出使用 k 最近邻（k-Nearest Neighbor，kNN）算法进行欺诈检测。针对传统 kNN 算法在搜索 k 近邻时耗时过多的问题，提出基于 KD 树结构的 kNN 算法。为解决经典 KD 树算法由于每次回溯都要回溯到根节点而导致查询效率低的问题，提出使用最佳桶优先（Best-Bin-First，BBF）算法进行 k 个近邻的查询。算法首先对待测数据集进行 PCA 降维，再构建 KD 树结构，最后使用 BBF 算法进行 k 近邻的查询。实验证明，提出的算法可及时有效地检测出欺骗行为。

关键词： 异常检测；k 最近邻；KD 树；BBF 算法；PCA 技术

The application of k-Nearest Neighbor algorithm based on improved KD tree in fraud detection

WU Jin'e，DUAN Qianqian
（School of Electronic and Electrical Engineering，Shanghai University of Engineering Science，Shanghai 201620，China）

【Abstract】 In the face of the problem that stores cheat consumers by brushing sales in Internet transactions，the k-Nearest Neighbor（kNN）algorithm is proposed to detect fraud. Aiming at the problem that traditional kNN algorithm spends too much time in searching k nearest neighbor，a kNN algorithm based on KD tree structure is proposed. In order to solve the problem of low query efficiency caused by the classical KD tree algorithm，Best-Bin-First（BBF）algorithm is used to query k nearest neighbors. First，PCA dimension reduction is performed for the measured data set，then KD tree structure is constructed，finally，BBF algorithm is used for k nearest neighbor query. Experimental results show that the proposed algorithm can detect the cheating behavior in time and effectively.

【Key words】 anomaly detection；k-Nearest Neighbor；KD tree；BBF algorithm；PCA technology

0 引 言

随着网络技术的发展，互联网交易现已是人们生活中必不可少的一部分，面对日益激烈的竞争，部分商家通过刷销量这种不良手段来博得消费者的信任与购买。面对这种状况，时下的欺诈检测方法主要通过区分正常数据与异常数据的差异来做出辨别，而异常检测方法也被公认为是有效的欺诈检测方法[1]。其中，基于 kNN 算法的异常检测技术即已广泛应用于各个领域中[2]。关于该算法在时间效率上的提升，主要分为 3 种：缩减数据集[3]、降低维度[4]、优化搜索空间[4]。KD 树算法则是索引空间优化中一种最经典、也最常用的算法，该方法可有效减少 k 个近邻的查询时间。

在异常检测中，准确率和时效性是评判算法优劣的重要指标。但是在高维数据中，KD 树需要回溯的节点数大大增加，这将会导致查询效率回退至传统 kNN 算法的蛮力搜索。因此改善高维数据的 k 近邻搜索方法，有助于快速发现并制止商家的欺骗行为，能有效地保护消费者的权益。

为了减少交易行为中欺诈行为的发生，学者们致力于研究领域适应且高效快速的欺诈检测算法。在信用卡欺诈检测方面，Mases 等人[5]分别将 BP 神经网络和贝叶斯信念网络模型应用于信用卡反欺诈场景中，实验结果表明贝叶斯信念网络具有更高的检测率。Liu 等人[6]针对金融欺诈提出了一种基于随机森林的检测模型，该模型利用特征选择，并使用真实数据集进行测试，实验验证了该算法具有较高的准确率。针对在线交易，在 Chang 等人[7]的工作中利用马尔可夫模型提出一种信誉评估机制，该算法利用粒子群优化算法的搜索机制快速捕捉电子商务系统中商家的不良行为，从而保护买家不受欺骗

基金项目： 国家重点研发计划（SQ2019YFB170208）。

作者简介： 吴金娥（1995-），女，硕士研究生，主要研究方向：数据异常检测研究；段倩倩（1986-），女，博士，讲师，主要研究方向：优化调度、数学建模及优化算法、非线性规划。

通讯作者： 段倩倩 Email：dqq1019@163.com

收稿日期： 2020-10-30

和其他恶意行为的侵扰。Wang 等人[8]提出了一套基于统计距离的监督式欺诈检测技术，该技术主要通过动态更新的阈值法检测异常，实验证明该技术可有效识别出异常。

1　相关技术

1.1　kNN 算法

kNN 算法[9]在分类和回归问题中都是常用算法之一，在回归问题中其判别异常的原理如下：首先基于某种距离度量计算每个待测数据点与其他数据点之间的距离；再分别找出待测数据点的 k 个最近邻之和，以和值作为待测数据点的异常值；最后通过比较异常值的大小判别异常数据点。

1.2　PCA 技术

主成分分析（Principal Component Analysis，PCA）技术又叫主分量分析技术[10]。该技术是一种常用的简化数据集的技术，主要通过利用降维的思想将多个指标转换成少数几个综合指标，现已广泛应用于模式识别、图像压缩以及异常检测等多个领域。

1.3　KD 树与 BBF 算法

经典的 KD 树（K-Dimensional Tree）算法是由 Bentley[11]在 1975 年提出的，KD 树是一种空间划分数据结构，在 kNN 算法上的应用包括建树和索引两个阶段。其中，索引阶段包括二分查找和回溯查找两个部分，前者确定查询路径，后者沿着查询路径逆向递归查询 k 个近邻至根节点。

在高维空间中，回溯次数的明显增多严重影响算法的效率。为解决这一问题，本文提出使用 BBF 算法[12]进行 k 个近邻的搜索。该算法通过建立优先队列并设置最高回溯次数与最大运行时间来提高算法执行的效率。在本文中，为高效快速地检测不良商家的欺诈行为，对传统的 kNN 算法进行改进，先使用 KD 树算法构建 KD 树，再使用 BBF 算法进行 k 个近邻的搜索。

2　基于改进 KD 树的 k 近邻算法

2.1　模型建立

穷举法是实现 kNN 算法最基础的方法。该方法需要计算每个数据点与数据集中所有数据点两两间的距离，当数据集较大时，该方法十分耗时，针对该问题，本文提出一种基于改进 KD 树的 kNN 算法。模型主要包括 3 部分，分别是：数据预处理模块、KD 树构建模块和 BBF 算法搜索模块。各部分功能简述如下。

（1）数据预处理模块：在该模块引入 PCA 技术对待测数据集做降维处理。

（2）KD 树构建模块：该模块使用类似于二叉树的结构存放数据，为 k 近邻的搜索做准备。该模块首先计算数据集 X 中 v 个维度的方差，再确定 split 域与根节点，最后通过与根节点 split 域值比较大小划分左右子树，依次递归至叶结点。

（3）BBF 算法搜索模块：BBF 算法的应用可有效减少回溯次数，加快 k 个最近邻的搜索。由于 KD 树搜索算法在高维数据集中的搜索效率因回溯次数的明显增加而显著下降，因此本文提出使用 BBF 算法替代 KD 树搜索算法，加快 k 个近邻的查找。该算法先通过节点自身携带的信息建立优先队列，每次回溯均从优先队列中取优先级最高的点 $t-node$，并比较该点与目标点的距离，对 k 近邻进行更新；然后比较在 split 域 p 维上该点与目标点值的大小，分别确定下一搜索节点与加入优先队列的节点，并比较下一搜索节点与目标点的距离，更新 k 近邻；最终递归搜索至优先队列为空或超出最大回溯次数，则结束搜索。

根据以上描述，算法模型建立如图 1 所示。

2.2　KD 树的建立

传统 kNN 算法依靠穷举法搜索 k 近邻，但这种方法在数据集较大时耗费过多时间，查询效率低下。而使用 KD 树结构可以在索引阶段避免大部分无关数据的查询，有效减少索引量，提高搜索效率。KD 树的构建过程是一个从根节点开始，自上而下逐级展开的递归过程。KD 树构建算法的详细描述如下。

算法 1　KD 树构建算法

输入　数据集 $X=\{X_1,X_2,\ldots,X_n\}$，数据点维度 v

输出　$KD-Tree$

1：for　i ← 1 to v do

2：　s_i ← 第 i 维数据的方差

3：end　for

4：spilt 域：p ← max(s_i) 所在维度

5：根节点：X_q ← p 维上的中位数所在的数据点

6：从数据集 X 中移除 X_q

7：if　$x_i^p < x_q^p$：

8：　X_i 进入 $l-tree$

9：else：

10：　X_i 进入 $r-tree$

11：重复上述步骤直至 X 为空

12：return　$KD-Tree$

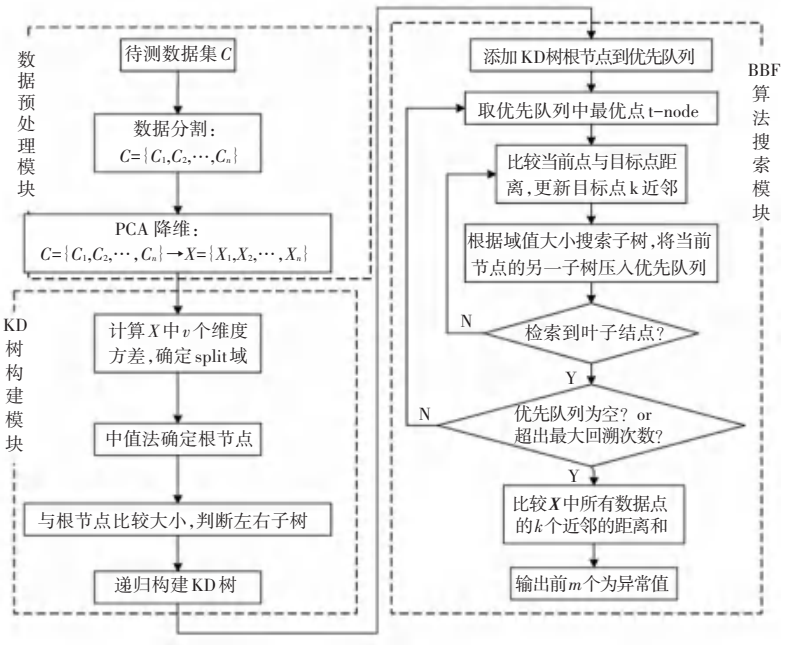

图 1 算法模型建立流程图

Fig. 1 Flow chart of algorithm model establishment

2.3 基于 BBF 算法的 k 近邻搜索

KD 树算法在搜索近邻时分为二分查找和回溯查找两个部分进行，可见回溯次数直接决定了算法的运行效率。而基于 KD 树搜索的 kNN 算法，每次搜索都需要回溯到根节点，出现了很多无关回溯，导致不必要的时间消耗。为提高 k 近邻的搜索效率，提出使用 BBF 算法，该算法提供一个优先队列，存储二分查找错过的节点，按照各节点所在的超平面到 A 的距离由小到大进行排序。对于指定查找点 A，其基于 BBF 算法的 k 近邻搜索见算法 2。

算法 2 基于 BBF 算法的 k 近邻搜索算法

输入 $KD-Tree$，A，最大回溯次数 H

输出 $N_k(A)$

1：$t-node \leftarrow prioritylist$ 出队最优节点

2：$c-node \leftarrow t-node$

3：if $dist(A,c-node) < \max(N_k(A))$

4： 更新 $N_k(A)$

5：if $A[split] > t-node[split]$

6： 将 $t-node$ 的 $l-tree$ 加入 $prioritylist$，$t-node$ 的 $r-tree$ 更新为 $c-node$

7：else

8： 将 $t-node$ 的 $r-tree$ 加入 $prioritylist$，更新 $t-node$ 的 $l-tree$ 为 $c-node$

9：判断是否检索到叶结点？否：执行步骤 3；是：执行下一步

10：$prioritylist$ 为空或超出 H？否：执行步骤 1；是：执行下一步

11：return $N_k(A)$

3 算法在欺诈数据集中的实验结果与分析

3.1 实验数据集

随着电子商务的兴起，部分商家为吸引消费者使用不良手段提升商品销量。本文针对该问题，选取来自于阿里巴巴天池大数据竞赛编号为 1629 的卖家交易数据，并将该数据以天为单位划分为 325 个集合，每个集合代表一个数据点。数据集的详细描述见表 1。

表 1 数据集介绍

Tab. 1 Data set introduction

数据集名称	数据类型	类型描述
原始数据集	正常交易数据	交易正常
	集中式刷销量数据	交易集中
	均衡式刷销量数据	交易均衡，总交易量高

由表 1 描述可知，数据集共分为 3 种类型的数据，其中集中式刷销量数据和均衡式刷销量数据描述了 2 种不同的刷销量模式。在检测时，使用正常交易数据混合 2 种刷销量数据构成待测数据集。

3.2 实验结果评价指标

混淆矩阵是异常检测算法实验结果最直观的体现，是其他评价算法优劣指标的根本。研究中会用到的混淆矩阵详见表 2。为更好地评价算法性能，本文选择 F_1 值、算法运行时间 T、ROC 曲线以及 PR 曲线作为实验结果的评价指标。

表 2　混淆矩阵

Tab. 2　Confusion matrix

类型名称	预测为异常	预测为正常
实际为异常	*TP*	*FN*
实际为正常	*FP*	*TN*

在异常检测中，F_1 很好地综合了精度(Precision，*Pre*)和召回率(Recall，*Re*)指标，是 *Pre* 和 *Re* 的加权调和平均。这里，*Pre* 指的是在预测为异常的集群中真实异常的占比；*Re* 表示预测为异常的集群占总真实异常的比例。如上各数学指标的计算公式可写为：

$$Pre = \frac{TP}{TP + FP}, \tag{1}$$

$$Re = \frac{TP}{TP + FP}, \tag{2}$$

$$F1 = 2 * \frac{Pre * Re}{Pre + Re}. \tag{3}$$

研究可知，F_1 值越大，则算法性能越好。

ROC 曲线被广泛应用于评估算法的可信度，描述了假阳率(False Positive Rate，*FPR*)和真阳率(True Positive Rate，*TPR*)之间的变化关系，其中 $FPR = \frac{FP}{FP + TN}$，而 *TPR* 等同于 *Re* 指标，该曲线越接近坐标系左上角，算法性能越好。*PR* 曲线分别以 *Re* 和 *Pre* 为横纵坐标，曲线越靠近图形右上角则算法性能越好。

3.3　实验分析

3.3.1　实验 1

实验 1 用于验证改进算法的时效性。传统 kNN 算法时间复杂度为 $O(n^2)$，而 KD 树结构可将时间复杂度降低为 $O(n\log n)$，利用 BBF 算法进行 k 近邻的查询，能进一步缩短搜索时间。在该组实验中，验证算法检测欺诈行为的时效性。待测数据集由正常交易数据和 20%的刷销量数据构成。其中，BBF-kNN 表示使用 KD 树结构存储数据，并使用 BBF 算法进行 k 近邻搜索的算法；PCA-BBF-kNN 表示使用 PCA 降维的 BBF-kNN 算法。实验结果记录见表 3。

表 3　算法时效性对比表

Tab. 3　Comparison table of algorithm timeliness

算法名称	集中式刷销量		均衡式刷销量	
	F_1 %	*T*/ms	F_1 /%	*T*/ms
kNN	**98.46**	2 538.91	76.00	2 526.54
BBF-kNN	98.15	976.67	**79.38**	624.11
PCA-BBF-kNN	97.84	**472.73**	69.54	**332.51**

由表 3 可知，集中式刷销量行为的最优 F_1 值可达到 98.46%，而均衡式刷销量模式的最优 F_1 值为 79.38%。可见均衡式刷销量模式更难检测，这是由于该模式下数据的分布比较相似增加了检测的难度。

在集中式刷销量模式下，kNN 算法的 F_1 值最高。BBF-kNN 算法在进行搜索时限制了最大回溯次数，一定程度上减少了数据量的对比；而 PCA-BBF-kNN 算法对 BBF-kNN 算法在数据预处理阶段进行了降维处理，信息量有所丢失，因此 F_1 值略低于 BBF-kNN 算法。尽管如此，kNN 算法下最优 F_1 值仅比 BBF-kNN 算法高 0.31%，比 PCA-BBF-kNN 算法高 0.62%；而其消耗的时间却是 BBF-kNN 算法的 2.6 倍，是 PCA-BBF-kNN 算法的 5.4 倍。在均衡式刷销量模式下，BBF-kNN 算法的 F_1 值最高，而 PCA-BBF-kNN 算法该值较低，这是由于在降维时丢失的信息较多，导致检测效果的下降。而算法的检测时间仍然是 PCA-BBF-kNN 算法最优。

综合上述分析可见，使用 KD 树结构合并 BBF 算法搜索 k 近邻的方法可有效检测出商家的欺诈行为。

3.3.2　实验 2

实验 2 为改进算法的可靠性验证。本组实验通过 *ROC* 曲线和 *PR* 曲线进一步验证算法的可靠性。仿真后得到，*ROC* 曲线如图 2 所示，*PR* 曲线如图 3 所示。

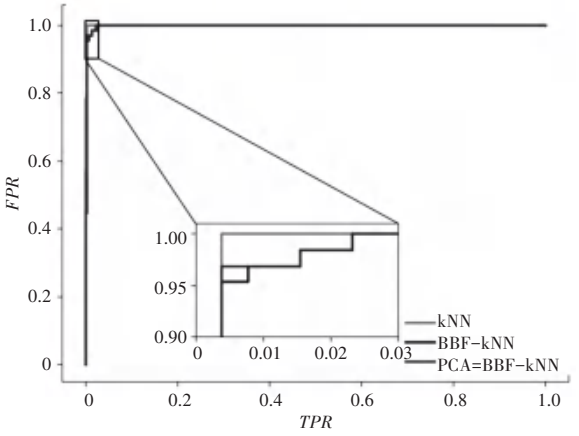

图 2　*ROC* 曲线

Fig. 2　*ROC* curve

对于 *ROC* 曲线而言，越接近点(0,1)，算法性能越好。由图 2 可见，实验中的 3 种算法都非常接近点(0,1)，因此本文提出的算法具有很好的可靠性。

算法的 *PR* 曲线越接近点(1,1),则性能越好,由图 3 可见,3 种算法的 *PR* 曲线表现良好,均接近点(1,1)。其中,kNN 算法最接近点(1,1),尽管如此,BBF-kNN 算法和 PCA-BBF-kNN 算法与 kNN 算法相差甚微。

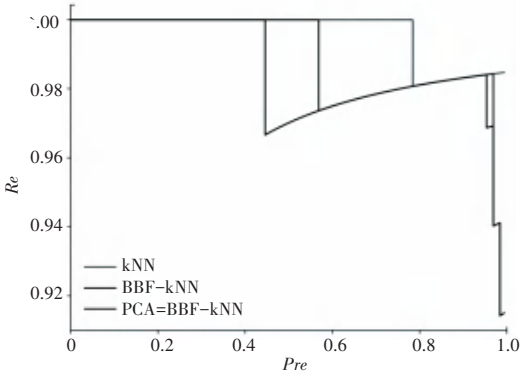

图 3 *PR* 曲线

Fig. 3 *PR* curve

4 结束语

本文针对互联网交易中存在的欺诈行为,提出了一种改进 KD 树的 k 近邻算法应用于其中. 对待测数据集进行 KD 树的构建,改善数据结构;再使用 BBF 算法搜索每个数据点的 k 近邻,减少回溯次数,缩短搜索时间;最后计算 k 个近邻之和,以此作为数据点的异常度,按照异常度的大小输出异常数据,从而判别商家的欺诈行为。实验验证了本文提出的算法可有效地减少执行时间,且检测准确率保持在较高的水平。

参考文献

[1] 高永昌. 医疗保险大数据中的欺诈检测关键问题研究[D]. 济南:山东大学,2018.

[2] MEHROTRA K G, MOHAN C K, HUANG H M. Anomaly detection principles and algorithms[M]. Switzerland: Springer International Publishing, 2017.

[3] VALERO-MAS J J, CASTELLANOS F J. Data reduction in the string space for efficient kNN classification through space partitioning[J]. Applied Sciences,2020, 10(10):3356.

[4] 江泽涛,周谭盛子,韩立尧. 基于感知哈希矩阵的最近邻入侵检测算法[J]. 电子学报,2019,47(7):1538-1546.

[5] MASES S, TUYLS K, VANSCHOENWINKEL B, et al. Credit card fraud bayesian and neural networks[C]// Proceedings of the 1st International Naiso Congress on Neuro Fuzzy Technologies. Havana, Cuba:Springer-Verlag, 2002: 16-19.

[6] LIU Chengwei, CHAN Yixiang, KAZMIL S H A, et al. Financial fraud detection model: Based on Random Forest[J]. International Journal of Economics & Finance, 2015, 7(7): 178-188.

[7] CHANG L, OUZROUT Y, NONGAILLARD A, et al. The reputation evaluation based on optimized Hidden Markov Model in e-commerce[J]. Mathematical Problems in Engineering, 2013, 2013: 391720.

[8] WANG Ruoyu, HU Xiaobo, SUN D, et al. Statistical detection of collective data fraud[C]//2020 IEEE International Conference on Multimedia and Expo(ICME). London, UK:IEEE, 2020:1-6.

[9] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[10] 杜子芳. 多元统计分析[M]. 北京:清华大学出版社,2016.

[11] BENTLEY J L. Multidimensional binary search trees used for associative searching[J]. Communications of the ACM, 1975, 18(9): 509-517.

[12] BEIS J S, LOWE D G. Shape indexing using approximate nearest-neighbour search in high dimensional spaces[C]// Proceeding of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). SAN JUAN, PR, USA:IEEE, 1997: 1000-1006.

(上接第 137 页)

5 结束语

本文针对渔业水质评价设计了基于 LBFGS 优化的神经网络模型,并对特征选取进行了深入讨论,进一步压缩水质评价模型,并使得准确率保持不变,因此更适合在前端嵌入式环境下运行。实验表明,本文设计模型能够有效提供准确的水质评价信息。

参考文献

[1] 胡朝莹,谢骏,余德光,等. 几种颜色空间在池塘水色图像识别中的稳定性研究[C]//大宗淡水鱼类产业技术可持续发展学术研讨会. 银川:中国水产科学研究院, 2009:17.

[2] 许新华. 基于 LM 神经网络水色图像识别技术的水质评价研究[J]. 科学技术创新, 2019(6):99-100.

[3] 王海英,曹晶,谢骏,等. 基于 L-M 神经网络优化算法的池塘水色判别系统的初步建立[J]. 渔业现代化, 2010,37(5):19-21,37.

[4] STRICKER M A, ORENGO M. Similarity of color images[C]// Proceedings of SPIE - The International Society for Optical Engineering 2420 . SAN JOSE, CA, USA:SPIE,1995:381-392.

[5] 张良均,王路,谭立云,等. Python 数据分析与挖掘实战[M]. 北京:机械工业出版社, 2016.