FM 集成模型在广告点击率预估中的应用

潘 博 张青川* 于重重 谢小兰
(北京工商大学计算机与信息工程学院 北京 100048)

摘 要 目前广告点击率预估所用的模型对于稀疏、类别分布不平衡的广告数据学习能力有限。针对这一问题,在数据分桶采样的基础上,提出利用因子分解机集成模型进行广告点击率的预估。利用迭代决策树算法提取的高层特征作为因子分解机的输入特征进行自动组合,发现特征间的相关性,解决数据稀疏和不均衡分类问题。在 Hadoop 大数据平台环境中对迭代决策树算法 + 因子分解机的融合模型进行并行式训练,可减少时间成本。通过单模型实验、采样实验、模型集成实验以及模型对比实验,确定了最佳采样比例,并验证了集成基于因子分解机的集成模型的有效性。

关键词 CTR 预估 FM 集成模型 Hadoop 大数据平台 互联网广告

中图分类号 TP311 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.01.018

THE APPLICATION OF INTEGRATION MODEL BASED ON FACTORIZATION MACHINE IN ADVERTISING CTR PREDICTION

Pan Bo Zhang Qingchuan* Yu Chongchong Xie Xiaolan
(*College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China*)

Abstract The current ad click rate estimation model has limited learning ability for sparse, unequal distribution of ad data. To solve this problem, this paper proposed an Integrated Model of Factorization Machine based on the respective data sampling to predict advertising CTR. The Gradient Boost Decision Tree Algorithm was used to extract the high-level features as the input features of the Factorization Machine to combine automatically, to find the correlation between the features, and to solve the problem of sparse data and imbalanced classification. In this paper, Hadoop was used to train the fusion model of Gradient Boost Decision Tree Algorithm + Factorization Machine in parallel to reduce the time cost. Through the single model experiment, model contrast experiment, the sampling experiment and the model integration experiment, the optimum sampling proportion was determined, and the validity of Integration Model based on Factorization Machine was verified.

Keywords CTR prediction Integration model of factorization machine Hadoop Internet advertising

0 引 言

随着广告计费方式的改变,广告点击率 CTR (Click-through Rate)估计在广告投放过程中占有越来越重要的地位。在预测效果的评价标准中,AUC (Area Under Curve)为 ROC(Receiver Operating Characteristics)曲线下的面积,可以反映分类器的平均性能,并对不同的 ROC 曲线进行比较,比起 ROC 曲线更能反映分类器效果。因此 AUC 值适合作为 CTR 预估模型的评价标准,一个完美分类器的 AUC 为 1.0,而随机猜测的 AUC 为 0.5。

随着机器学习技术的发展,这一技术被逐渐采用到广告点击率预估中。基于机器学习的 CTR 预估方法一般可以分为三种:线性机器学习模型、非线性机器学习模型,以及融合模型。线性机器学习模型具有简

收稿日期:2017 - 03 - 19。北京市自然科学基金重点项目(KZ201410011014);北京市教委科研计划面上项目(KM201510011009);北京市教委科研计划面上项目(KM201510011010)。**潘博**,硕士生,主研领域:数据挖掘,机器这习。**张青川**,讲师。**于重重**,教授。**谢小兰**,硕士生。

单易实现、易扩展、易在线更新的特点，可以处理超大规模的数据，因此在产业界应用较为广泛。文献[1]首次将 CTR 预估问题由概率估计问题转成回归问题。该文提出用逻辑回归 LR(Logistic Regression)来解决 CTR 预估问题，将具体广告、环境抽象成特征，用特征来达到泛化的目的，从而对广告点击率进行预测。文献[2]提出了基于 L-BFGS 的 OWLQN(Orthant Wise Limitedmemory Quasi Newton)优化算法，解决了逻辑回归损失函数在 L1 正则化非连续可导的问题。这个方法的优点在于：这是一种 Batch Learning 的方法，可以收敛到全局最优解，且收敛速度快；损失函数中使用了 L1 正则化，避免过拟合的同时输出稀疏模型。

线性机器学习模型虽然简单易扩展，但是表达能力有限，不能学习特征间的非线性关系，即使构造出复杂特征 + 轻量模型的形式也不一定对线性模型有效，而且费时费力。因此非线性模型也被研究用来预估广告 CTR，形成简单特征 + 复杂模型的形式。文献[3]首次提出使用点击日志计算搜索结果的点击率，结合搜索引擎查询日志和用户点击日志，自动优化搜索引擎的检索质量。通过分析用户在当前返回的排序结果中点击链接的日志，使用支持向量机算法最大化 Kendall 相关系数，从而达到排序结果接近最佳排序的目的。文献[4]提出了一种在线贝叶斯概率回归模型 OBPR(Online Bayesian Probability Regression)用于在搜索广告情景下对广告 CTR 进行预测。

为了进一步提高点击率预估的效果，近些年来融合模型也得到了一些尝试。文献[5]针对 Facebook 的社交广告点击率预估研究，提出了迭代决策树算法 GBDT(Gradient Boost Decision Tree) + LR 的融合模型。该方法结合了 GBDT 非线性模型能拟合非线性特征的特点，以及 LR 线性模型具有很好的扩展性以及模型训练速度快的特点，融合之后的 AUC 比 GBDT 和 LR 单模型均高出 3%。文献[6]提出了人工神经网络 + 决策树(ANN + MatrixNet)融合模型。ANN 用来拟合 ID 类离散特征，输出点击概率。MatrixNet 为 Yandex 公司版本的 GBDT，用来拟合连续特征和 ANN 模型输出的特征。该融合模型有效地融合了 CTR 预估中的分类特征和连续特征，实验效果比 LR、MatrixNet 和 ANN 单模型要好。

广告数据呈现的高维稀疏性和特征之间存在着高度非线性关联的特点，虽然上述融合模型能同时具有非线性模型与线性模型的优点，但是 MatrixNet、GBDT 和 LR 等模型仍然无法有效学习特征间的非线性关系，所以对于高维稀疏、类别不平衡的广告数据无法建立准确的 CTR 预估模型。本文将针对已有的数据量大、高维稀疏、类别不平衡的广告数据集，重点研究 FM 集成模型以使广告点击率预估具有更高的效率和更好的模型精度。

1 FM 集成模型原理

1.1 问题描述

本文研究的广告点击率预估问题可以描述为：输入一个用户查询和其他信息(如性别、年龄、地域、兴趣爱好等)，经过点击率预估系统计算输出每一则广告的点击概率，即广告的预估点击率。图 1 中的黄色模块描述了一个广告点击率预估系统的构建流程。

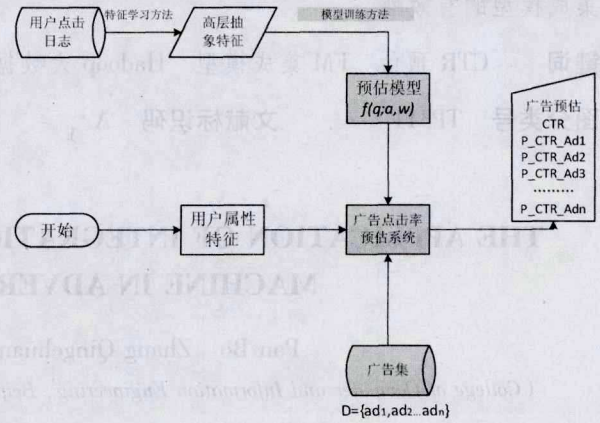

图 1 点击率预估系统工作流程

针对高维稀疏、类别不平衡的广告数据建立广告点击率预估模型是本文研究重点。本文利用因子分解机 FM(Factorization Machine)模型自动拟合特征间的交互，处理高维稀疏问题；同时提出了 FM 集成模型，解决了分类不均衡问题。本文结合文献[7]中 GBDT 的高层特征提取方法形成 GBDT + FM 融合模型，利用大数据处理平台 Hadoop 的分布式系统对该模型进行并行式训练，减少模型对海量广告数据的训练时间。

1.2 集成学习概述

集成学习用多重或多个弱分类器结合为一个强分类器，从而达到提升分类效果的目的。文献[8]对解决不均衡问题的集成模型进行了分类，集成学习主要是基于 Bagging[9] 和 Boosting[10] 两种基本思想。Bagging 通过自举汇聚原始训练集来训练不同的分类器。也就是从原始数据集随机采样 N 次得到 N 个数据集，通常这 N 个数据集和原始数据集保持同样的大小，从而通过重采样过程获得数据多样性。N 个数据集建好后，将某种学习算法分别作用于每个数据集，得到 N 个分类器。当对新数据进行分类时，应用这 N 个分类器进行分类，选择分类器投票结果中最多的类别为最后的分类结果。

Boosting 与 Bagging 很类似,使用的多个分类器的类型都是一样的。但是在 Boosting 中,不同的分类器是通过串行训练获得,每个新分类器都根据已训练出的分类器的性能进行训练,通过集中关注被已有分类器错分的那些数据来获得新的分类器。Boosting 分类的最终结果是根据所有基分类器的权重加权求和得到的,该权重代表的是分类器在上一轮迭代中的成功度。AdaBoost 算法[14]已成为目前最流行的 Boosting 算法,该算法的效率与 Freund 改进算法很接近,却可以非常容易地应用到实际问题中。

本文数据集中的未被点击的广告数(反例数目)比被点击的广告数(正例数目)多出 25 倍,显然属于非均衡分类问题。为了解决非均衡分类问题,本文采用多重数据抽样方法,构造基于数据集多重抽样的集成分类器,同时采用了 AUC 作为模型评价指标。

1.3　FM 集成模型

FM 是由 Steffen Rendle[11-12]提出的一种基于矩阵分解的机器学习算法。其最大的特点是对于稀疏的数据具有很好的学习能力。本文涉及广告 CTR 预估,其数据同样具有高维稀疏的特点,因此采用 FM 模型作为基分类器,结合了 Bagging 和 Boosting 的集成思想进行集成学习。由于本文中广告原始数据集正负比例相差较大,类别分布不均衡。为了使输入到单个分类器的样本均衡,本文将正负样本分开存储,分别进行采样。其中正例可全采样或过采样,负例样本欠采样,得到的 N 个数据集进行转换后分别输入 FM 模型进行单独训练,每个 FM 单模型会有自己的输出和模型 AUC 指标。各个模型的输出通过自身 AUC 加权求和形成最终的 FM 集成分类器,其算法流程图见图 2。

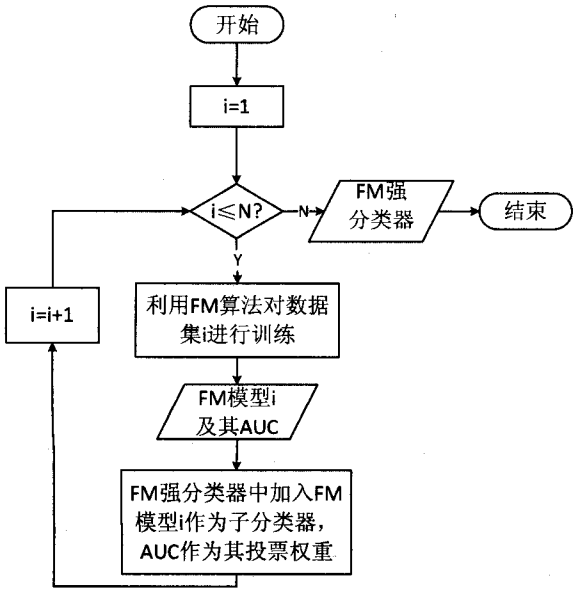

图 2　FM 集成算法流程图

一般 Bagging 所用的基分类器为不稳定算法,如决策树,集成才会有较大的效果。此处稍微不同于 Bagging 思想的地方在于,一方面为了达到样本类别均衡,另一方面尽量不丢失原始数据模式。FM 模型为稳定算法,此处集成是为了拟合不同分桶中的负例样本。通常广告数据量很大,对 N 个数据集进行串行式训练会消耗大量时间。针对这个问题,本文基于大数据平台 Hadoop 中的 MapReduce 离线计算框架进行并行化训练,从而提高效率,其模型见图 3。

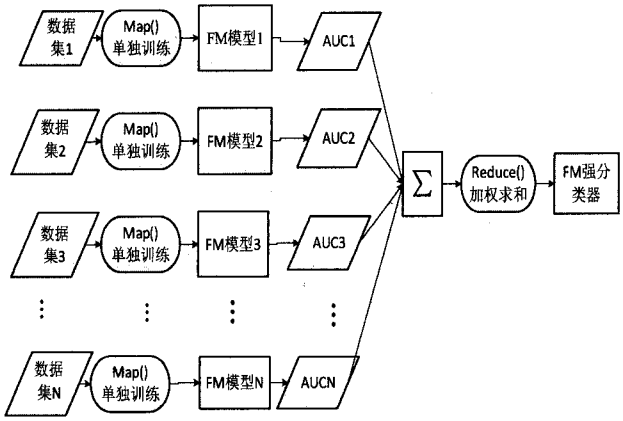

图 3　FM 集成算法的并行化训练模型图

负样本是基于 Hive 分桶采样的,如将负样本分为 N 桶,将第 1 桶负列样本和正例样本混合作为训练集 1,将第 2 桶负列样本和正例样本混合作为训练集 2,以此类推,将第 N 桶负列样本和正例样本混合作为训练集 N。分桶的大小需要经过实验来确定。通过 Map 对每个数据集同时进行单独训练出对应 FM 模型,然后通过 Reduce 对各个模型的输出通过自身 AUC 加权求和,形成最终的 FM 集成分类器。该模型既实现了 Bagging 和 Boosting 的集成思想的结合,又降低了训练时间成本,从而实现了精度与效率的双赢。

2　广告点击率预估模型实验

2.1　实验数据集与实验环境介绍

本文采用 3 个实验数据集:① 腾讯多媒体展示广告数据集。② SIGKDD Cup2012 track2[15]提供的广告点击日志数据。③ 国内最大的定向广告联盟——百度联盟的点击日志。三个数据集包含会话日志、用户信息、广告信息三方面内容。

本文采用文献[7]中基于 GBDT 特征提取方法,不仅能提升模型的效果,而且可以减少特征工程的成本和时间,因此可利用该算法分别提取出其中的 ID 类特征与 GBDT 类特征(高层特征)。针对类标签信息,从三个数据集中抽出相同数量训练集、测试集,信息如

表 1 所示。

表 1 训练集、测试集类标签信息

类标签	训练集	测试集
正例(1)	746 056	97 530
负例(0)	19 141 712	2 634 174

实验平台为基于 Hadoop 的大数据处理平台，用到的组件包括 HDFS 分布式文件系统、MapReduce 离线计算框架、YARN 分布式资源管理系统、Hive 分布式数据仓库和 Spark 的 MLlib 机器学习算法库。实验中集群规模为 1 个 Master 节点，3 个 Slave 节点，集群网络拓扑信息如图 4 所示。

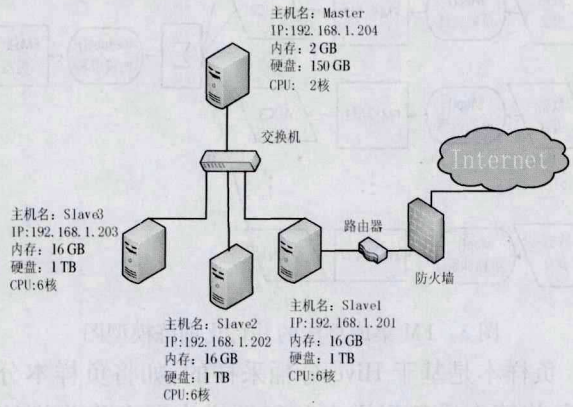

图 4 集群网络拓扑

2.2 实验总体框架

实验总体框架如图 5 所示。

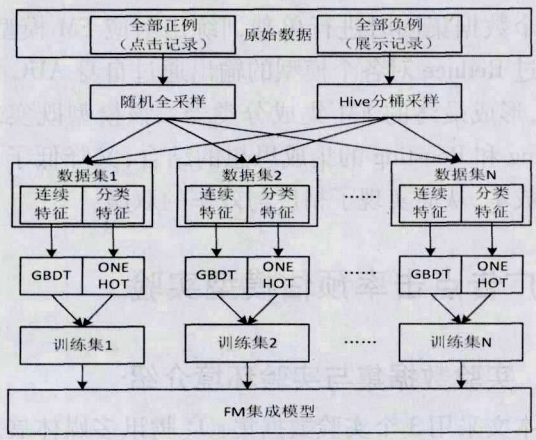

图 5 FM 集成模型实验总体框架

原始数据按正例(点击记录)和负例(展示记录)分别存放到不同的表。对于正例进行随机全采样，即采样样本的数据量等于原始正例的数量；对负例进行分桶采样，第 1 桶与正例样本组成数据集 1，第 2 桶与正例样本组成数据集 2，以此类推，第 N 桶与正例样本组成数据集 N。采样得到的各个数据集由连续特征和分类特征组成，对于连续特征训练 GBDT 模型提取高层二值特征向量；对于分类特征进行 ONE HOT 独热编码，形成高维稀疏特征。将两种特征组合在一起形成最终的 N 个训练集。这些训练集用来学习最终的 FM 集成模型。

2.3 实验过程

2.3.1 单模型实验

本文采用的三个广告数据集均为稀疏数据。本文利用 GBDT 模型的多维特征提取方法提取出的高层特征，并联合独热编码后的 ID 类特征一起输入到 CTR 预估模型。为了验证 FM 模型比 LR 模型能更好处理高维稀疏数据，对两者进行了对比实验。实验结果如表 2 所示。

表 2 FM 与 LR 在三个数据集上的 AUC

数据集	FM:AUC	LR:AUC
腾讯广告	0.76	0.691 0
SIGKDDCUP2012track2	0.73	0.682 2
百度联盟广告	0.70	0.650 0

由表可以看出，FM 模型在三个稀疏数据集都表现出比 LR 模型更好的预估性能。FM 因子分解机基于因子分解，能拟合特征之间的相关性，可以有效地滤除高维稀疏数据中的“噪声”，提取出权重较高的特征。

2.3.2 采样实验

当训练数据类别分布极度不平衡时，严重影响模型的预测精度。因此在本文中，当负样本多余正样本两个数据量级时，需采取合理的采样策略使训练样本尽量平衡且不丢失有效信息。本文实验了在三个数据集上的多种正负样本比例时 FM 模型的效果，AUC 取平均值，结果如图 6 所示。

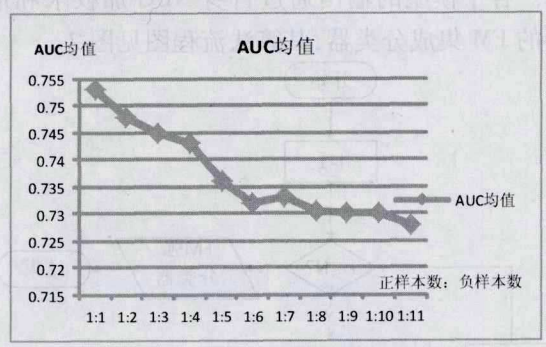

图 6 不同正负样本比例下 FM 模型的 AUC 指标

由图 6 可见，AUC 随负样本的比重增多呈现出下降的趋势。当正负比例 1:1时 AUC 最大，为 0.752 4。除了以上的正负比之外，本实验也测试了全量数据下(正负比为 1:25)FM 模型效果，其 AUC 值为 0.605 4。当对正样本重采样，负样本欠采样(正负比分别为2:2，2:3，2:4)时，AUC 停留在 0.5 左右，即模型没有预测效果。故本文将采用正负样本采样比为 1:1的数据集

进行单模型训练。

2.3.3 模型集成实验

对于不均衡分类问题,必须经过采样使训练样本尽量平衡。但采样之后的数据有可能丢失大量的有效信息,即便确定了合理的采样比例,还是会有信息的丢失,所以提出了 FM 集成模型,将分桶采样后的数据分别输入 FM 单模型训练,然后将单模型输出按 AUC 加权求和,得到最终的 FM 集成模型输出。实验的输入数据为 ID 类特征,同 2.3.2 节一样在三个数据集上取 AUC 均值为实验结果,如图 7 所示。其中横坐标 1 代表只有一个单模型,2 代表第一个单模型与第二个单模型集成,3 代表第三个单模型与前两个模型集成,依次类推,25 代表第 25 个单模型与前面 24 个单模型集成。

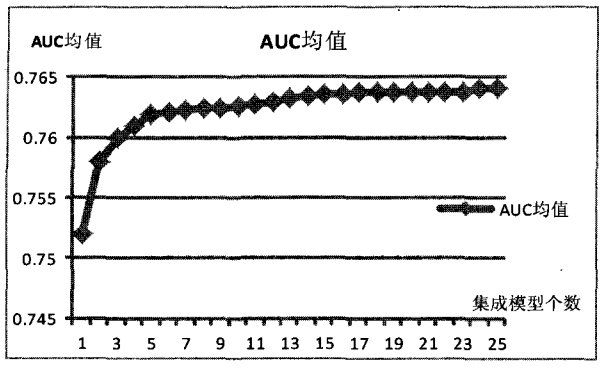

图 7　FM 集成模型指标与集成模型个数关系图

由图 7 可知,前 5 个模型累计集成的时候变化较大,后面随着单模型的加入 AUC 仍然在上升,但上升相对缓慢。FM 集成模型的 AUC 比 FM 单模型 AUC 提高 0.01 以上,效果较明显。

2.3.4 模型对比验证

实验将 ID 类特征和 GBDT 类特征输入分别输入到 LR 模型、FM 模型和 FM 集成模型进行 AUC 比较,进一步验证 FM 模型对于其他输入特征的有效性。采用的训练数据条数为 1 511 992 条,正负样本比为1:1。与文献[7]相同,ID 类特征通过独热编码后的维数为 3 963 094 维,连续类特征通过训练好的 GBDT 模型后得到的高层特征维数为 3 840($2^7 \times 30$)维。实验结果如表 3 所示。

表 3　FM 模型与 FM 集成模型对比实验结果

数据集	模型	特征	AUC
腾讯广告	LR	ID 类	0.690 000
	FM	ID 类	0.752 200
	FM 集成	ID 类	0.763 00
	LR	ID 类 + GBDT 类	0.720 000
	FM	ID 类 + GBDT 类	0.780 152
	FM 集成	ID 类 + GBDT 类	0.791 000
SIGKDD-Cup2012track2	LR	ID 类	0.650 000
	FM	ID 类	0.683 000
	FM 集成	ID 类	0.743 000
	LR	ID 类 + GBDT 类	0.700 000
	FM	ID 类 + GBDT 类	0.750 400
	FM 集成	ID 类 + GBDT 类	0.763 600
百度联盟广告	LR	ID 类	0.634 000
	FM	ID 类	0.663 400
	FM 集成	ID 类	0.703 000
	LR	ID 类 + GBDT 类	0.680 000
	FM	ID 类 + GBDT 类	0.730 170
	FM 集成	ID 类 + GBDT 类	0.742 000

由表 3 可知,在三个数据集上,无论输入的是 ID 类特征还是 ID 类特征 + GBDT 类特征,FM 集成模型 AUC 指标均为最高,比 FM 模型高 0.01 以上,比 LR 模型高出 0.07 以上;LR 模型的 AUC 最低,其表达能力有限,不能学习特征间的非线性关系。显然,相比其他模型,FM 集成模型具有更好学习能力。

3 结　语

本文首先对集成学习进行了概述,涉及到解决不均衡问题的集成模型的分类,介绍了 Bagging 和 Boosting 两种基本思想;然后提出了本文的 FM 集成模型架构;最后通过采样实验、模型集成实验、模型对比验证,实验验证了 FM 集成模型的有效性。

类别不平衡问题在互联网广告点击率预估当中是常见问题,本文提出了 FM 集成模型,结合 GBDT 高层特征提取方法,形成 GBDT + FM 的融合模型,并用 Hadoop 实现了并行化。该模型在减少建模人工成本和时间成本的同时,也能有效提高精度。

参 考 文 献

[1] McMahan H B, Holt G, Sculley D, et al. Ad click prediction: a view from the trenches[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1222 - 1230.

[2] Shan L, Lin L, Sun C, et al. Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization [J]. Electronic Commerce Research and Applications, 2016, 16:30 - 42.

(下转第 148 页)

4 结 语

本文通过对 ZigBee 网络层路由算法节能问题的研究，提出了一种新的路由算法，通过对节点进行区域划分，并确定路由过程中的最短区域顺序，以达到在路由过程中对 RREQ 请求分组进行定向的目的，减少网络中冗余的分组。同时结合节点的深度选择合适的路由算法，当路由节点深度较低时，选择使用 AODVjr 算法，当节点深度较高时，选择使用 Cluster-Tree 簇树路由算法，以此避免高深度时大规模转发 RREQ 请求分组的问题，从而在保证路由算法发现路径较佳的同时，降低整体网络的能耗。该算法平衡了 AODVjr 算法与 Cluster-Tree 算法的优缺点，较为有效地降低网络整体能耗，提高了 ZigBee 网络整体性能。

参 考 文 献

[1] 钱志鸿，王义君. 面向物联网的无线传感器网络综述[J]. 电子与信息学报，2013，35(1)：215－227.

[2] 谢川. 基于 ZigBee 的 AODVjr 算法研究[J]. 计算机工程，2011，37(10)：87－89.

[3] 谢川. ZigBee 中改进的 Cluster-Tree 路由算法[J]. 计算机工程，2011，37(7)：115－117.

[4] 徐沛成，胡国荣. 改进的 ZigBee 网络路由算法[J]. 计算机工程与设计，2013，34(9)：3019－3023.

[5] 钱志鸿，朱爽，王雪. 基于分簇机制的 ZigBee 混合路由能量优化算法[J]. 计算机学报，2013，36(3)：485－493.

[6] Mu J. A directional broadcasting algorithm for routing discovery in ZigBee networks[J]. EURASIP Journal on Wireless Communications and Networking，2014，2014(1)：94.

[7] Chakeres I D，Klein-Berndt L. AODVjr，AODV simplified [J]. Acm Sigmobile Mobile Computing & Communications Review，2002，6(3)：100－101.

[8] Kim T，Kim S H，Yang J，et al. Neighbor Table Based Shortcut Tree Routing in ZigBee Wireless Networks[J]. IEEE Transactions on Parallel & Distributed Systems，2014，25(3)：706－716.

[9] 朱旭，牛存良，白晓丽. 改进的 ZigBee 树路由算法[J]. 计算机工程与应用，2016，52(5)：114－118.

[10] Xie H F，Zeng F，Zhang G Q，et al. Simulation Research on Routing Protocols in ZigBee Network[C]//Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation. Atlantis Press，2016.

(上接第 111 页)

[3] Agarwal D，Agrawal R，Khanna R，et al. Estimating rates of rare events with multiple hierarchies through scalable log-linear models[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM，2010：213－222.

[4] Dembczynski K，Kotlowski W，Weiss D. Predicting ads click-through rate with decision rules[C]//Workshop on targeting and ranking in online advertising，WWW'08. 2008.

[5] Shen S，Hu B，Chen W，et al. Personalized click model through collaborative filtering[C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM，2012：323－332.

[6] Agarwal D，Chen B C，Elango P. Spatio-temporal models for estimating click-through rate[C]//Proceedings of the 18th international conference on World wide web. ACM，2009：21－30.

[7] 田嫦丽，张珣，潘博，等. 互联网广告点击率预估模型中特征提取方法的研究与实现[J]. 计算机应用研究，2017，34(2)：334－338.

[8] Galar M，Fernandez A，Barrenechea E，et al. A review on ensembles for the class imbalance problem：bagging-，boosting-，and hybrid-based approaches[J]. Systems，Man，and Cybernetics，Part C：Applications and Reviews，IEEE Transactions on，2012，42(4)：463－484.

[9] Erdal H，İlhami Karahanoğlub. Bagging Ensemble Models for Bank Profitability：An Emprical Research on Turkish Development and Investment Banks[J]. Applied Soft Computing，2016，49：861－867.

[10] Rok Blagus，Lara Lusa. Gradient boosting for high-dimensional prediction of rare events[J]. Computational Statistics and Data Analysis，2016.

[11] Rendle S. Factorization machines[C]//Data Mining (ICDM)，2010 IEEE 10th International Conference on. IEEE，2010：995－1000.

[12] Rendle S. Social network and click-through prediction with factorization machines[C]//KDD-Cup Workshop，2012.

[13] Richardson M，Dominowska E，Ragno R. Predicting clicks：estimating the click-through rate for new ads[C]//Proceedings of the 16th international conference on World Wide Web. ACM，2007：521－530.

[14] Pouria Ramzi，Farhad Samadzadegan，Peter Reinartz. An AdaBoost Ensemble Classifier System for Classifying Hyperspectral Data[J]. Photogrammetrie，Fernerkundung，Geoinformation，2014(1)：27－39.

[15] SIGKDD12 Cup[OL]. http://www.kddcup2012.org/c/kddcup2012-track2.