号影响, 使得小区边缘的区域接收到的干扰比小区内部的区域接收到的干扰大. 所以单一的指标不能很好的反映实际问题. 在这种情况下, 对于质差小区的判断不能只依靠覆盖率这一指标, 还应当考虑到信号点位置的影响. 因此在本研究中, 引入了信号点距离的因素, 对于距离基站较远的信号点, 可以根据实际情况, 将判断的阈值设置一个合适的比近区信号点判断阈值低的数值.

ComputerTime	HandsetTime	Longitude	Latitude	SINR	PCI	RSRP
2016/7/12 11:28	28:28.0			25.1	89	-86.93
2016/7/12 11:28	28:28.0			22.1	89	-87.87
2016/7/12 11:28	28:28.0			22.1	89	-87.87
2016/7/12 11:28	28:27.5			25.6	89	-85.87
2016/7/12 11:28	28:28.1			23.4	89	-87.25
2016/7/12 11:28	28:27.5			24.6	89	-86.06
2016/7/12 11:28	28:28.1	107.9685567	21.54902348	23.1	89	-86.93
28:46.0	28:27.6	107.9685568	21.54902357	22.7	89	-87
28:46.5	28:28.1	107.968557	21.54902405	23.3	89	-86.37
28:46.2	28:28.1	107.968557	21.5490241	23.3	89	-86.37
28:46.2	28:27.8	107.9685571	21.54902414	23.1	89	-86.56
28:46.5	28:28.1	107.9685573	21.54902452	23.4	89	-86.25
28:46.2	28:27.9	107.9685573	21.54902462	24.9	89	-86.31
28:46.5	28:28.1	107.9685575	21.54902495	23.6	89	-86.18
28:46.3	28:28.1	107.9685575	21.54902505	23.6	89	-86.18
28:46.5	28:28.0	107.9685575	21.5490251	23	89	-87.68
28:46.3	28:28.0	107.9685576	21.54902514	23	89	-87.68
28:46.5	28:28.2	107.9685578	21.54902557	22.4	89	-87.68

图 1　部分原始数据图

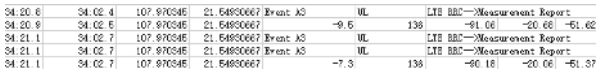

图 2　乱码数据图

Longitude	Latitude	SINR	RSRP	PCI
107.9685567	21.54902348	23.1	-86.93	89
107.9685568	21.54902357	22.7	-87	89
107.968557	21.54902405	23.3	-86.37	89
107.968557	21.5490241	23.3	-86.37	89
107.9685571	21.54902414	23.1	-86.56	89
107.9685573	21.54902452	23.4	-86.25	89
107.9685573	21.54902462	24.9	-86.31	89
107.9685575	21.54902495	23.6	-86.18	89
107.9685575	21.54902505	23.6	-86.18	89
107.9685575	21.5490251	23	-87.68	89
107.9685576	21.54902514	23	-87.68	89

图 3　预处理后的部分数据图

本文提出了基于距离因素的四维特征, 分别为 SINR 近区好点比例、SINR 远区好点比例、RSRP 近区好点比例和 RSRP 远区好点比例. SINR 近区好点比例为近区 SINR>3 dBm 的信号点的比例; SINR 远区好点比例为远区 SINR>0 dBm 的信号点的比例; RSRP 近区好点比例为近区 RSRP>–90 dBm 的信号点的比例; RSRP 远区好点比例为远区 RSRP>–100 dBm 的信号点的比例. 其中, 将距离基站最近的信号点和最远的信号点的距离的平均值作为阈值, 大于阈值的区域为远区, 小于或等于阈值的区域为近区. 为了分析特征的可分性, 更加直观的观察特征, 将 RSRP 近区好点比例作为 x 轴, RSRP 远区好点比例作为 y 轴, 得到图 4; 将 SINR 近区好点比例作为 x 轴, SINR 远区好点比例作为 y 轴, 得到图 5.

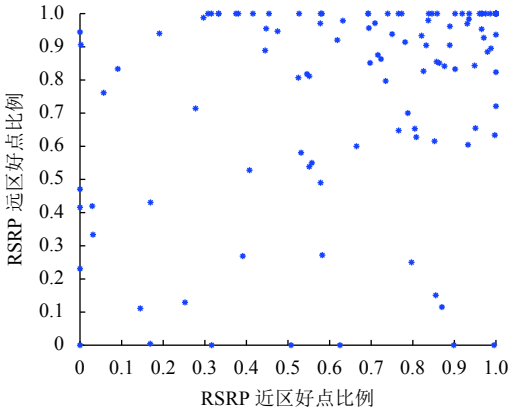

图 4　RSRP 好点比例图

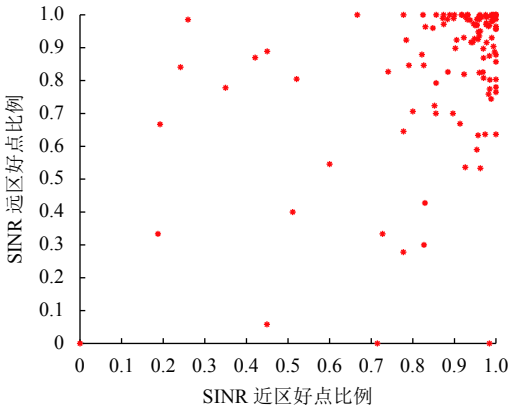

图 5　SINR 好点比例图

从图 4 和图 5 可以看出, 数据点集中在 0.7–1 之间, 说明 117 个小区里, 覆盖率较好的小区或者干扰较弱的小区占比较高. 优良小区集中分布, 呈现较好的集聚现象. 由此可得, 特征能将优良小区和质差小区区分出来, 可以用分类器进行分类.

1.4 数据标定

目前对于质差小区的检测, 很大程度上依据的是网优工作人员的经验. 传统的路测, 需要网优人员结合多个质量指标的统计数据, 利用路测分析软件对小区进行判断和评估, 存在正确率不够、效率低下等问题.

为了提高基于路测数据对质差小区检测的效率, 更为准确的判断小区的优劣情况, 采用聚类算法结合人工标注的方法进行标定. 本研究采用的聚类算法为 k 均值聚类 (k-means clustering algorithm, k-means) 算法. 该方法是最为常用的一种无监督算法. 首先随机选择 k 个点作为质心, k 的选值需要人为设定. 再计算数据集中的每一个点离质心的欧式距离或者余弦距离等, 将其分配到距其最近的质心所在的簇. 之后每个簇的质心更新为这个簇中所有点的平均值, 直到满足终止条件. 这种算法简单快速容易实现, 能够体现数据在几何和统计学上的意义[10].

先利用 k 均值聚类算法, 将 k 值设定为 2, 即将所有的小区划分为两类, 简单分析后将优良小区标注为 1, 质差小区标注为 0. 网优人员结合聚类结果, 在路测分析软件上进行分析, 将结果进一步细化, 得到最终分类结果. 如图 6 所示, 在 PCI 为 115 的小区中有近一半的信号点的 RSRP 值低于阈值, 因此小区覆盖下的信号强度不理想, 应为质差小区. 同时查看聚类结果, PCI 为 115 的小区的标定值为 0, 则聚类结果正确, 小区判定为质差小区. 对于判断不一致的小区, 交给另一位人员进行判断, 得到最终结果, 减少了误判率, 提高了工作的效率和判断的准确率.

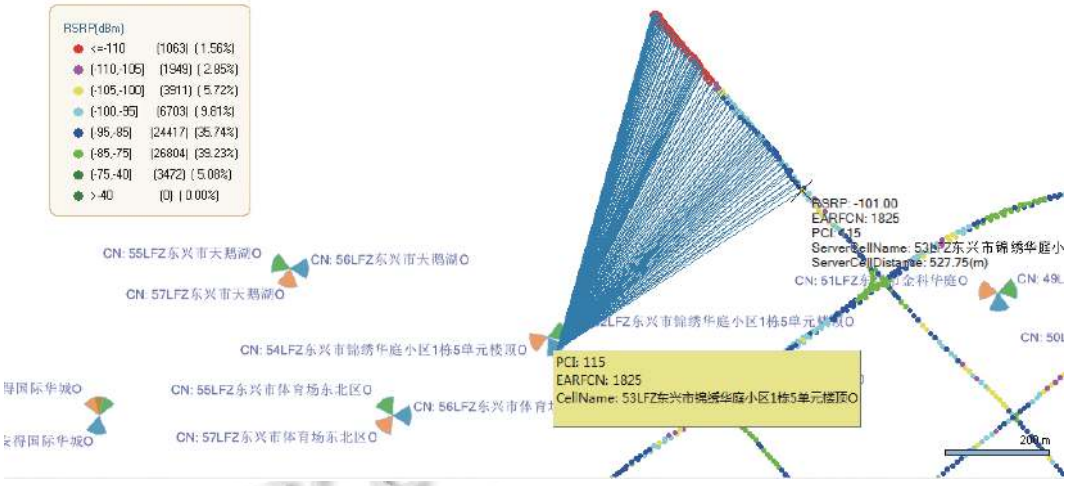

图 6 路测软件分析图

2 分类器选择

2.1 选择标准

分类器根据学习的方式主要分为无监督学习分类器和有监督学习分类器. 无监督学习的分类器可以利用未标记的数据, 找到其中的隐藏结构, 根据样本之间的相似性进行分类; 监督学习的分类器依据标签, 在分类好的数据基础上判断一个新的数据所属的类别.

选择分类器时, 既要考虑分类器本身的特性, 又要考虑各式数据集在训练时带来的影响. 无线网络的 LTE 小区的路测数据经过数据处理、特征提取、数据标注后, 产生 117×4 的特征矩阵和标签, 将其输入到分类器中, 可以看出, 训练数据为一个小样本, 维度较高的数据集, 应当选择属于适用于小样本的分类器. 结合实际, 实际的网优工作, 分类器应该具有较好的可解释性, 可以在工作中, 提供较好的指导性. 考虑到实际的工程应用, 分类器的计算复杂度, 要选择快速且资源消耗小的算法.

2.2 算法概述

2.2.1 逻辑回归算法

逻辑回归 (logistics regression) 算法是监督学习的

一种常用算法, 主要解决二分类问题. 假设训练集 $D=\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(m)},y^{(m)})\}$, 其中特征 $X^{(i)}\in R^n$, 类的标记 $y^{(i)}\in\{0,1\}$, 假设函数如下:

$$h_\theta(x)=\frac{1}{1+\exp(-\theta^{\mathrm{T}}X)} \tag{1}$$

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}Cost(h_\theta(x^i),y^i) \tag{2}$$

$$Cost(h_\theta(x),y)=-\log(h_\theta(x)),\ \text{if}\ y=1 \tag{3}$$

$$Cost(h_\theta(x),y)=-\log(1-h_\theta(x)),\ \text{if}\ y=0 \tag{4}$$

其中, θ为模型参数, $J(\theta)$为损失函数. 逻辑回归算法的最终目的就是最小化损失函数$J(\theta)$. 这种算法运行速度快, 简单易于理解, 容易更新模型, 但是对数据和场景的适应能力有一定的局限性.

2.2.2 支持向量机算法

支持向量机 (Support Vector Machine, SVM) 算法, 是基于统计学习理论的一种监督机器学习的方法. 支持向量机可以找到一个最优分类超平面, 这个超平面能够使其两侧的空白区域最大化, 而且不失分类的精度[11]. 它在小样本数据上能够得到较好的结果, 而且具有优秀的泛化能力. 但是运用在大数据集上会出现训练时间过长和准确率不够的问题.

2.2.3 决策树算法

决策树算法属于监督学习, 可以分为分类树和回归树. 分类树可以基于不同的条件分割数据集. 首先根据信息增益或者信息增益率来寻找最优特征, 然后根据特征中的最优值将数据集分成两个子数据集, 之后重复以上操作, 直到满足终止条件. 信息增益和信息增益率的公式为:

$$Gain(D,a)=Ent(D)-\sum_{v=1}^{m}p_vEnt(D^v) \tag{5}$$

$$Gain\_ratio(D,a)=\frac{Gain(D,a)}{-\sum_{v=1}^{m}p_v\mathrm{Log}_2p_v} \tag{6}$$

其中, a 是特征, v 是其中的一个分类, p_v是 v 分类占特征 a 总个数的比例, D^v为根据 v 分类进行划分之后的数据集, $Ent(D^v)$是划分后数据集的信息熵, $Gain(D,a)$表示根据特征 a 划分之后的信息增益, $Gain_ratio(D,a)$表示信息增益率. 这种算法速度快, 准确率高, 可生成易理解的规则, 但是对于样本数据量不一致的数据比较敏感, 容易忽略掉属性之间的相关性.

2.2.4 k 最近邻算法

k 最近邻 (k-Nearest Neighbor, kNN) 分类算法是一种广泛应用的监督学习算法. k 近邻算法遇到一个未知类别的新样本时, 根据一些已知类别的样本, 可以找到 k 个距离最小的邻居样本. 新样本就属于类别中含有这些邻居数量最多的类. 这种算法理论基础成熟, 准确度高, 但对于大数据集来说, 计算量大, 所需内存多, 会造成运行时间过长等问题.

综上所述, 选择逻辑回归分类器、支持向量机分类器、决策树分类器和 k 近邻分类器这 4 种较为简单的算法作为质差小区的检测的算法, 并通过结果对比分析, 得到最适合的分类器.

3 实验与结果分析

3.1 实验设计

本文提出的基于机器学习的质差小区检测方法的整体流程图如图 7 所示.

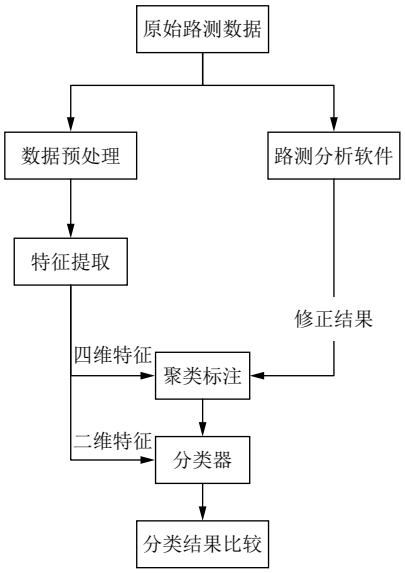

图 7 整体流程图

整个过程主要分为 3 个部分: 特征提取、数据标注和分类结果比较. 其中特征提取可以为分类器提供训练数据基础; 数据标注将质差小区的检测问题, 转化成机器学习中分类问题; 分类器结果比较是将两维特征 (即 RSRP 采样率和 SINR 采样率) 与基于距离的四维特征分别输入到每一种分类器中进行分类, 将其结果进行比较; 若四维特征得到较高的分类准确率, 则将其准确率和运行时间综合比较, 最终选择出分类效果

最好的分类器.

为了最小化模型结构风险, 本实验采用 10 折交叉验证对性能进行评估, 将数据集平均分成 10 份, 轮流将其中的 9 份做训练, 1 份做验证, 并将 10 次结果的均值作为对算法精度的估计, 得到每种算法较为合理的准确率.

3.2 特征的比较

为了验证四维特征的可行性, 分别使用逻辑回归分类器、支持向量机分类器、决策树分类器和 k 近邻分类器 4 种分类器对提取的四维特征和二维特征分别进行分类比较. 二维特征和四维特征在不同分类器下的准确率结果见表 1.

表 1 二维特征和四维特征在不同分类器下的准确率

分类器	两维特征	四维特征
逻辑回归	0.829	0.915
决策树	0.803	0.906
支持向量机	0.846	0.932
k 近邻	0.769	0.863

由表 1 可以看出, 四维特征比二维特征在每一种分类器中的分类准确率都高 10% 左右, 由此可得, 四维特征具有更高的分类准确率, 证明了四维特征的可行性, 说明了基于距离的四维特征在基于机器学习的质差小区的检测中具有一定的实际意义.

3.3 分类器的选择

选取逻辑回归分类器、支持向量机分类器、决策树分类器和 k 近邻分类器 4 种分类器, 得到了四维特征在该 4 种分类器中的结果如表 2 所示.

表 2 四维特征在不同分类器下的准确率和运行时间

分类器	准确率	运行时间 (s)
逻辑回归	0.915	4.7427
决策树	0.906	4.901
支持向量机	0.932	1.2945
k 近邻	0.863	1.1964

由表 2 可以看出, 在这 4 种分类器, 支持向量机分类器, 表现出了更加优异的分类性能 (准确率高且运行时间短). 且四维特征在该分类器下得到的混淆矩阵和 ROC 曲线如图 8 和图 9 所示.

由图 8 可得, 人工标注质差小区为 43 个, 其中有 38 个被正确预测为质差小区, 有 5 个被错误的预测为优良小区; 人工标注优良小区为 74 个, 有 71 个被成功预测为优良小区, 3 个被错误预测为质差小区. 可以看出, 支持向量机分类器对于四维特征的分类效果良好, 错误分类在可接受的范围内.

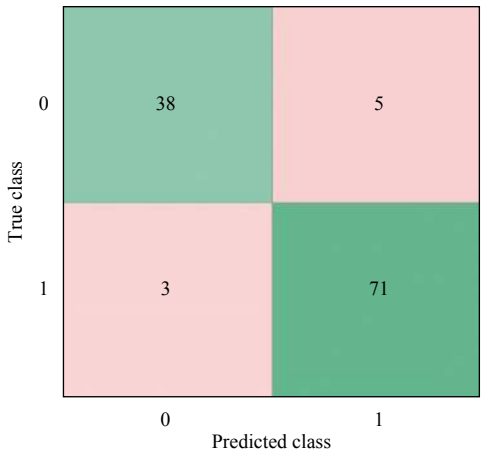

图 8 混淆矩阵

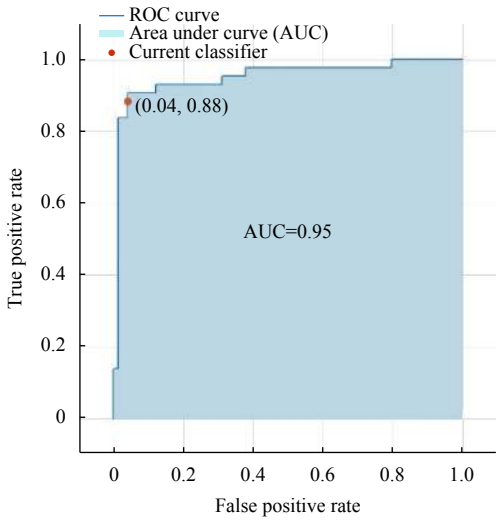

图 9 ROC 曲线

ROC 曲线的横轴代表负正类率特异度, 纵轴代表真正类率灵敏度, 通过曲线可以很容易的查到任意界限值时的分类能力.AUC 为 ROC 曲线下的面积, 它作为数值可以直观的评判分类器的好坏, 取值在 0.1 到 1 之间, 由图 9 可以看出, 支持向量机分类器的 AUC 为 0.95, 说明支持向量机分类器性能比较好.

通过实验可以发现, 对于常用的 4 种分类器, 本文提出的基于距离的四维特征相比传统的两维特征均获得了更高的准确率, 说明了将距离因素引入质差小区的检测能得到更加准确的分类结果, 其中在支持向量机中得到了最好的结果.

4 结论

本研究将距离因素引入到传统的路测数据中, 得到了基于距离的四维特征. 分析比较了二维特征与四维特征在逻辑回归分类器、支持向量机分类器、决策树分类器和 k 近邻 4 种分类器中的效果, 并分析了四维特征在 4 种分类器中的分类准确率和运行时间. 根据以上研究可以得出:

(1) 使用四维特征与二维特征进行机器学习的分类检测比较, 四维特征能够获得较好的区分结果.

(2) 对比逻辑回归分类器、支持向量机分类器、决策树分类器和 k 近邻分类器 4 种分类器, 在二维特征和四维特征中, 支持向量机分类器均获得了最好的分类效果.

所以, 将距离因素引入到对路测数据进行质差小区检测能够得到更好的结果, 解决了单一指标在质差小区检测中准确度不够的问题, 在路测数据中为机器学习在质差小区检测中的应用提供了理论依据, 具有一定的现实意义.

参考文献

1 果敢, 于力, 魏然. 无线网络优化的路测. 电信技术, 2005, (1): 20–22. [doi: 10.3969/j.issn.1000-1247.2005.01.007]

2 王西点, 王磊, 龙泉, 等. 人工智能及其在网络优化运维中的应用. 电信工程技术与标准化, 2018, 31(7): 81–86. [doi: 10.3969/j.issn.1008-5599.2018.07.019]

3 张喆. 基于 K-means 的 LTE 宏站小区场景聚类策略. 通信技术, 2019, 52(3): 668–673. [doi: 10.3969/j.issn.1002-0802.2019.03.026]

4 Kibria MG, Nguyen K, Villardi GP, *et al*. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. IEEE Access, 2018, 6: 32328–32338. [doi: 10.1109/access.2018.2837692]

5 周鹏. 数据挖掘在通信网络优化中的应用研究[硕士学位论文]. 南京: 南京邮电大学, 2017.

6 曾雨桐. 数据挖掘算法在网络优化话务量和差小区挖掘中的应用[硕士学位论文]. 南京: 南京邮电大学, 2016.

7 王希. 基于概率神经网络 (PNN) 的 LTE 质差小区分析方法. 数字通信世界, 2017, (2): 89–90, 80. [doi: 10.3969/J.ISSN.1672-7274.2017.02.029]

8 林世明, 高志斌, 高凤连, 等. 基于路测的 TD-LTE 网络优化分析. 现代电子技术, 2015, 38(9): 12–15. [doi: 10.3969/j.issn.1004-373X.2015.09.003]

9 龙青良, 张磊. 基于用户感知的 LTE 网络优化关键问题研究. 邮电设计技术, 2014, (10): 14–20. [doi: 10.3969/j.issn.1007-3043.2014.10.004]

10 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述. 计算机应用研究, 2007, 24(1): 10–13. [doi: 10.3969/j.issn.1001-3695.2007.01.003]

11 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述. 电子科技大学学报, 2011, 40(1): 1–10. [doi: 10.3969/j.issn.1001-0548.2011.01.001]