一种基于SMOTE的不均衡样本KNN分类方法*

林泳昌，朱晓姝**
(玉林师范学院计算机科学与工程学院，广西玉林　537000)

摘要:针对在数据样本不均衡时，K 近邻(K-nearest Neighbor，KNN)方法的预测结果会偏向样本数占优类的问题，本文提出了一种基于合成少数类过采样方法(SMOTE)的KNN不均衡样本分类优化方法(KSID)。该方法过程为:首先使用SMOTE方法将不均衡的训练集均衡化，并训练逻辑回归模型;然后使用逻辑回归模型对训练集进行预测，获取预测为正样本的数据，通过使用SMOTE方法均衡化该正样本，并训练KNN模型;最后把测试集放入该结合逻辑回归方法的KNN模型进行预测，得到最终的预测结果。围绕6个不均衡数据集，将KSID与逻辑回归、KNN和支持向量机(SVM)决策树等方法进行对比实验，结果表明，KSID方法在准确率、查全率、查准率、$F1$ 值这4个性能指标上均优于其他3种方法。通过引入SMOTE，KSID方法克服了KNN模型遇到样本不均衡数据集时，产生分类偏向的问题，为进一步研究KNN方法的优化和应用提供参考。
关键词:不均衡样本　KNN　SMOTE　KSID　逻辑回归　分类

中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1005-9164(2020)03-0276-08
DOI:10.13656/j.cnki.gxkx.20200707.001

0　引言

K 近邻(K-nearest Neighbor，KNN)方法由Cover和Hart于1968年提出，使用节点的邻居节点信息构建最近邻图来做分类，是机器学习中常用、简单、易实现的二分类方法。当前，很多数据集具有不均衡性，比如信用卡欺诈、用户违约预测数据集等。这导致KNN方法面临一个挑战:当数据样本不均衡时，KNN方法的预测结果会偏向于样本数占优类。为了提高KNN方法的运行效率、分类效果等性能，很多学者对KNN方法进行了优化。比如沈焱萍等[1]提出基于元优化的KNN方法。王轶凡[2]提出数据时效性的高效KNN方法。王志华等[3]提出基于改进K-modes聚类的KNN方法。张万桢等[4]提出环形过滤器的K值自适应KNN方法。余鹰等[5]使用变精度粗糙集上下近似概念，增强了KNN方法的鲁棒性。樊存佳等[6]采用改进的K-Medoids聚类方法裁剪对KNN分类贡献小的训练样本，提高KNN的分类精度。罗贤锋等[7]使用K-Medoids方法对文本训练集进行聚类，解决了传统KNN方法在文本训练集过大时存在速度慢的问题。针对不均衡

*国家自然科学基金项目(61762087)，广西自然科学基金项目(2018JJA170175)和大学生创新创业计划项目(201810606014)资助。
【作者简介】
林泳昌(1998—)，男，本科，主要从事大数据分析研究。
【**通信作者】
朱晓姝(1973—)，女，教授，硕士生导师，主要从事生物信息、大数据分析、分布式网络计算等研究，E-mail:xszhu@ylu.edu.cn。
【引用本文】
林泳昌，朱晓姝.一种基于SMOTE的不均衡样本KNN分类方法[J].广西科学，2020，27(3):276-283.
LIN Y C，ZHU X S. A SMOTE-based KNN Classification Method for Unbalanced Samples [J]. Guangxi Sciences，2020，27(3):276-283.

数据集,KNN 方法预测结果会偏向于样本数占优类的问题,本文将 Synthetic Minority Oversampling Technique(SMOTE)[8] 和逻辑回归引入 KNN 方法中,提出了一种基于 SMOTE 的 KNN 不均衡样本分类优化方法(A Modified KNN Algorithm based on SMOTE for Classification Imbalanced Data,KSID)。使用 SMOTE 改进逻辑回归方法或改进 KNN 方法,虽然可以提高不均衡数据集的查全率,但也会很明显地降低模型查准率;特别是在大数据集上,使用 SMOTE 会大大增加 KNN 方法的时间复杂度。因此,本文先使用 SMOTE 对训练集做上采样,并使用训练过的逻辑回归对该均衡的训练集预测分类,再利用 SMOTE 和 KNN 对预测为正样本的数据集做上采样并预测分类,从而训练得到新的 KNN 模型,以有效地解决 SMOTE 会降低模型查准率和增加时间复杂度的问题。

1 材料与方法

1.1 逻辑回归

逻辑回归通过构建损失函数,并进行优化,迭代,求解出最优的模型参数,最后得到逻辑回归分类模型。其方法如下[9]:

步骤 1:随机初始化 W 和 b,利用公式(1)计算预测标签。

$$Z=W^TX+b\ , \tag{1}$$

其中,Z 为预测结果,W 为权重矩阵,X 为特征矩阵,b 为偏移量。

步骤 2:利用公式(2)计算模型的损失函数。

$$J(a,b)=\frac{1}{m}\sum_{i=1}^{m}L(a^{(i)},y^{(i)})\ , \tag{2}$$

其中,m 为样本数,$a^{(i)}$ 为样本 i 的真实标签,$y^{(i)}$ 为样本 i 的预测标签。

步骤 3:利用公式(3)(4)计算 W、b 的梯度并更新。

$$W=W-\alpha\frac{\partial(J(W,b))}{\partial W}\ , \tag{3}$$

$$b=b-\alpha\frac{\partial(J(W,b))}{\partial b}\ , \tag{4}$$

其中,α 为学习率。

步骤 4:迭代步骤 1 到步骤 3,直到最小化损失函数 $J(a,b)$。

1.2 KNN 方法

KNN 是一个简单而经典的机器学习分类方法,通过度量待分类样本和已知类别样本之间的距离(一般使用欧氏距离),对样本进行分类。其方法如下[10]:

步骤 1:根据公式(5)计算样本点到所有样本点的欧氏距离(d)。

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2}\ 。 \tag{5}$$

步骤 2:根据欧式距离的大小对样本进行排序(一般是升序)。

步骤 3:选取前 K 个距离最近的邻居样本点。

步骤 4:统计 K 个最近的邻居样本点分别属于每个类别的个数。

步骤 5:将 K 个邻居样本点里出现频率最高的类别,作为该样本点的预测类别。

可以看出,当数据样本不均衡时,KNN 方法预测结果会偏向于样本数占优类。

1.3 SMOTE

SMOTE 是常用于样本不均衡数据集的采样方法[8]。其思路是通过合成一些少数类样本,增加少数类样本个数,使得样本均衡。SMOTE 的生成策略:对于每个少数类样本 x,从它的最近邻样本中随机选一个样本 y,然后在 x、y 属性的欧氏距离之间随机合成新的少数类样本,其工作原理如图 1 所示。

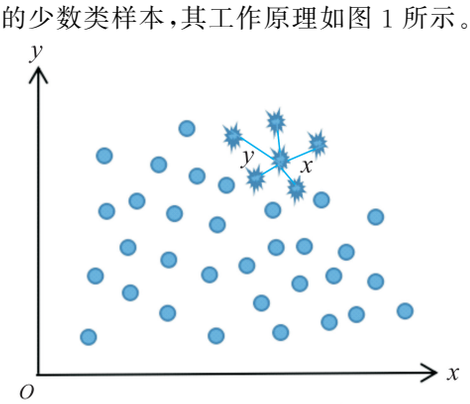

图 1 SMOTE 工作原理

Fig. 1 Working principle of the SMOTE method

SMOTE 的步骤如下[8]:

步骤 1:在少数样本集中,对于少数类的样本 x,利用公式(5)计算 x 到其他所有样本的欧氏距离,得到 K 个最近邻点。

步骤 2:通过设置采样比例,得到采样倍率 N,并对每一个少数类样本 x,从其 K 近邻中随机选取近邻样本 y。

步骤 3:对于近邻样本 y 和样本 x,通过计算公式(6),构建新样本。

$$L = x + \text{rand}(0,1) \times |x - y|, \tag{6}$$

其中,x 表示少数样本,y 表示 x 的最近邻样本。

1.4 KSID 方法

将 SMOTE 和逻辑回归引入 KNN 方法中,提出了一种基于 SMOTE 的 KNN 不均衡样本分类优化方法(KSID)。针对数据样本不均衡对 KNN 二分类器分类效果影响的问题,使用 SMOTE 进行采样,增加少数样本的数量,并消除数据样本不均衡对 KNN 分类效果的影响。KSID 方法描述如下:

步骤 1:使用 SMOTE 将不均衡训练集均衡化。将不均衡数据集按照一定的比例划分为训练集和测试集。使用 SMOTE 对训练集中的每个少数样本,计算其与其他少数样本的欧氏距离,得到 K 个近邻样本点。通过设置采样比例,随机选取少数类样本,对每一个选出的少数类样本,利用公式(6)构建新样本,一直迭代到训练集的正负样本均衡终止。

步骤 2:构建逻辑回归模型。将经过步骤 1 采样后的训练集放入逻辑回归模型,训练其正则化惩罚项、学习率等参数。并使用训练好的逻辑回归模型,对训练集进行预测。

步骤 3:生成 KNN 模型的训练集并利用 SMOTE 均衡化。将步骤 2 中预测为正样本的数据集,作为 KNN 模型的训练集,并使用 SMOTE 进行采样,使得训练集的正负样本均衡。

步骤 4:构建 KNN 模型。使用步骤 3 输出的训练集,训练 KNN 模型的参数 K,并得到模型结果。

步骤 5:预测测试集标签。把测试集放入步骤 2 构建的逻辑回归模型预测,将预测为正样本的数据放入步骤 4 构建的 KNN 模型预测,并得到预测结果;最后将 KNN 模型预测标签与逻辑回归模型预测为负样本的标签合并,得到测试集的预测标签。

算法 1 给出了 KSID 算法的形式描述。

算法 1:KSID 算法

Begin

输入:训练数据 x_train,训练标签 y_train,近邻数 K,倍率 N

输出:新样本类别 y_pred

0:将 x_train、y_train、K、N 传入 SMOTE 得到数据集 newdata

1:将 newdata 传入逻辑回归方法得到预测结果 y_pred

2:y_true=-y_train[y_pred==1]

3:x_true=-x_trian[:,y_pred==1]

4:将 x_true、y_true、K、N 传入 SMOTE 得到数据集 newdata1

5:将 newdata1 传入逻辑回归方法得到预测结果 y_true_pred

6:y_pred[y_pred==0].append(y_true_pred) #合并结果

7:输出结果 y_pred

End

KSID 方法虽然将训练集通过 SMOTE 采样,可以消除数据不均衡性,但是 SMOTE 对训练集均衡化后,产生合成的正样本影响分类性能。针对这个问题,将逻辑回归预测的正样本,继续使用 KNN 进行预测,进而提高分类性能。且 KSID 方法先使用逻辑回归模型预测出大量正确的负样本,再使用 KNN 预测少量的正样本,可以有效降低模型计算复杂度,减少模型计算时间。

1.5 实验数据来源

本文实验数据集为信用卡欺诈数据集、员工离职数据集、企业诚信数据集、广告点击预测数据集、用户违约数据集、疝气病症预测病马是否死亡数据集,前 5 个数据集均来源于 DC 竞赛网(https://www.dcjingsai.com/),第 6 个数据集来源于百度百科网。数据集的具体描述如表 1 所示,其中样本不均衡率计算公式如下[11]:

$$\text{样本不均衡率} = \frac{\text{正样本数}}{\text{负样本数}}。 \tag{7}$$

1.6 实验环境

使用 notebook 编译软件、Python 3.6 语言编程,分别使用 sklearn 里的 KNN 方法、逻辑回归方法和 imblearn 包中的 SMOTE。计算机硬件配置为 8 GB 内存、64 位操作系统、i5-6300HQ 处理器。

1.7 实验参数设置

训练集和测试集划分比例为 7∶3,随机种子设置为 0;逻辑回归设置 fit_intercep 为 True;SMOTE 中的 K 取值为 5,SMOTE 采样倍率设置为 5;支持向量机(SVM)决策树方法[12]最大深度设置为 4;逻辑回归模型和 KNN 模型都使用 3 折交叉验证。

1.8 实验方案设计

将本文提出的 KSID 方法分别与逻辑回归方法、KNN 方法、SVM 决策树方法相比较。基于 6 个数据集分别测试这 4 种方法的准确率、精确度、查全率、$F1$ 值、运行时间等性能指标,并分析实验结果。

表 1 实验数据集

Table 1 List of the data sets

数据集 Data set	样本 Sample	变量 Variable	正样本数 Positive samples	负样本数 Negative samples	不均率 Uneven rate (%)
信用卡 Credit card	284 807	29	492	284 315	0.17
员工离职 Labor turnover	1 100	35	178	922	19.30
企业诚信 Enterprise integrity	14 366	7	941	13 425	7.00
广告点击 Click advertising	1 001 650	39	198 787	802 863	24.76
用户违约 User default	700	12	183	517	35.40
疝气病症 Colic symptoms	299	21	121	178	68.00

2 结果与分析

2.1 评价指标

在机器学习的分类方法中,常常使用准确率来衡量模型的分类效果。但对于不均衡数据集,还需要使用查全率、查准率、$F1$ 值等指标评价模型性能。这 4 种评价指标都是基于表 2 的正负样本混淆矩阵。

其中 TP 和 TN 分别表示正确分类的正类和负类的样本数量;FN 和 FP 分别表示错误分类的正类和负类的样本数量。

表 2 混淆矩阵

Table 2 Confusion matrix of samples

分类 Classification	预测为正类 Predicted to be positive	预测为负类 Predicted to be negative
正类 Positive	TP	FP
负类 Negative	FN	TN

1)准确率(Accuracy)

准确率是衡量模型的总体分类效果的指标,准确率越高模型分类效果越好[13]:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\times 100\%。\quad (8)$$

2)精确率(Precision)

精确率是模型预测的正样本在真实正样本中所占的比例[11]:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}。\quad (9)$$

3)查全率(Recall)

查全率指有多少个正样本被模型正确预测为正样本[11]:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}。\quad (10)$$

4)$F1$ 值($F1$ score)

$F1$ 值是精确率和查全率的调和值,其结果更接近两者较小的值[11]:

$$F1=2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}。\quad (11)$$

2.2 准确率测试实验

对 6 个数据集进行准确率测试。从表 3 的测试结果可以看出,KSID 方法在 6 个数据集上准确率的均值基本大于其他 3 种方法,即 KSID 的分类性优于其他 3 种方法,特别是员工离职数据集,比 KNN 方法提高了 4.2%。这是因为 KSID 使用逻辑回归进行第一次分类,再使用 KNN 对第一次分类预测为正样本的数据进行第二次分类,进而提高了模型的准确率。但对于样本量和特征维度较大的广告点击数据集,KSID 方法使用 SMOTE 采样产生较多的伪样本,影响模型对原始样本的分类准确率,使得 KSID 方法的分类准确率低于 SVM 决策树方法。

2.3 精确率测试实验

对 6 个数据集进行精确率测试,其测试结果如表 4 所示。

表 3　准确率测试结果

Table 3　Results of accuracy test results

数据集 Data set	逻辑回归 Logistic regression	*K* 近邻 KNN	SVM 决策树 SVM decision-making tree	KSID
信用卡 Credit card	0.999 2	0.999 4	0.999 5	0.999 5
员工离职 Labor turnover	0.861 0	0.849 0	0.848 0	0.891 0
企业诚信 Enterprise integrity	0.881 0	0.920 0	0.930 0	0.931 0
广告点击 Click advertising	0.681 0	0.771 0	0.802 0	0.780 0
用户违约 User default	0.814 0	0.781 0	0.705 0	0.819 0
疝气病症 Colic symptoms	0.722 0	0.722 0	0.744 0	0.778 0
均值 Mean	0.826 0	0.840 0	0.838 0	0.866 0

表 4　精确率测试结果

Table 4　Results of precision test results

数据集 Data set	逻辑回归 Logistic regression	*K* 近邻 KNN	SVM 决策树 SVM decision-making tree	KSID
信用卡 Credit card	0.884	0.728	0.898	0.938
员工离职 Labor turnover	0.526	0.456	0.462	0.700
企业诚信 Enterprise integrity	0.000	0.323	0.429	0.358
广告点击 Click advertising	0.000	0.305	0.671	0.362
用户违约 User default	0.750	0.643	0.400	0.653
疝气病症 Colic symptoms	0.790	0.790	0.818	0.914
均值 Mean	0.492	0.541	0.613	0.654

从表 4 可以看出,KSID 方法在 6 个数据集上精确率的均值基本大于其他 3 种方法,即 KSID 模型对正样本预测更准确,如信用卡欺诈数据比 KNN 方法提升了 21%、员工离职数据比 KNN 方法提高了 24.4%等。这是因为 KSID 使用 KNN 模型对正样本进行二次分类,进而提高了模型的精确率。但对于企业诚信和广告点击等数据集,由于 KSID 使用逻辑回归进行第一次分类,而逻辑回归无法识别正样本(其精确率为 0),影响了模型的分类精确率,使得 KSID 方法的分类精确率低于 SVM 决策树方法。

2.4　查全率测试实验

对 6 个数据集进行查全率测试,从表 5 的测试结果可以看出,KSID 方法在 6 个数据集上查全率的均值基本大于其他 3 种方法,即 KSID 模型对正样本的召回数量更多,如在信用卡欺诈数据集中,KSID 模型的查全率比逻辑回归模型高 23.8%、比 KNN 模型高 13.6%等。这是因为 KSID 使用 SMOTE 将不均衡数据集均衡化,改进 KNN 方法对不均衡数据集分类偏向的缺点,进而提高了模型的查全率。但对于员工离职和广告点击等数据集,由于 KSID 使用 KNN 对正样本进行二次分类,而 KNN 对正样本查全率不高,影响了模型的分类查全率,使得 KSID 方法的分类查全率低于逻辑回归和 SVM 决策树方法。

表 5 查全率测试结果

Table 5 Resules of recall test results

数据集 Data set	逻辑回归 Logistic regression	*K* 近邻 KNN	SVM 决策树 SVM decision-making tree	KSID
信用卡 Credit card	0.619	0.721	0.782	0.857
员工离职 Labor turnover	0.436	0.286	0.250	0.417
企业诚信 Enterprise integrity	0.000	0.209	0.010	0.670
广告点击 Click advertising	0.000	0.190	0.604	0.462
用户违约 User default	0.444	0.333	0.300	0.593
疝气病症 Colic symptoms	0.776	0.552	0.775	0.776
均值 Mean	0.379	0.382	0.454	0.629

2.5 *F*1 值测试实验

对 6 种数据集进行 *F*1 值测试,从表 6 的测试结果可以看出,KSID 方法在 6 个数据集上 *F*1 值的均值基本大于其他 3 种方法,如用户违约数据集 KSID 模型的 *F*1 值比逻辑回归模型高 6.4%、比 KNN 模型高 18.3%、比 SVM 决策树模型高 28.2%等。但对于广告点击数据集,由于 KSID 使用逻辑回归和 KNN 进行第一、第二次分类,而 KNN 和逻辑回归对正样本识别率不高(其 *F*1 值较低),导致影响了模型的分类 *F*1 值,使得 KSID 方法的分类 *F*1 值低于 SVM 决策树方法。

表 6 *F*1 值的测试结果

Table 6 Resules of *F*1 score test results

数据集 Data set	逻辑回归 Logistic regression	*K* 近邻 KNN	SVM 决策树 SVM decision-making tree	KSID
信用卡 Credit card	0.728	0.725	0.836	0.896
员工离职 Labor turnover	0.477	0.356	0.325	0.523
企业诚信 Enterprise integrity	0.000	0.264	0.020	0.467
广告点击 Click advertising	0.000	0.249	0.636	0.406
用户违约 User default	0.558	0.439	0.340	0.622
疝气病症 Colic symptoms	0.783	0.650	0.796	0.840
均值 Mean	0.424	0.447	0.492	0.626

2.6 运行时间测试实验

对 6 种数据集进行运行时间测试,从表 7 的测试结果可以看出,对于小数据集,KNN 方法的运行时间小,但对于大数据集(比如用卡欺诈数据集),KSID 方法时间复杂度明显比 KNN 方法小得多,如信用卡欺诈数据集 KSID 运行时间比 KNN 快 266.545 s 等。这是因为 KSID 方法先使用逻辑回归方法预测了量大的负样本,再使用 KNN 方法预测量少的正样本,进而降低了方法时间复杂度。此外,逻辑回归方法总的运行时间最少,但它在 4 个预测性能指标上都是最低。KSID 既能获得最好的预测性能,也能极大地降低运行时间。

表 7　运行时间测试结果

Table 7　Run time of test results

数据集 Data set	逻辑回归 Logistic regression	*K* 近邻 KNN	SVM 决策树 SVM decision-making tree	KSID
信用卡 Credit card	6.734	280.113	16.058	13.568
员工离职 Labor turnover	0.138	0.020	0.073	0.209
企业诚信 Enterprise integrity	0.146	0.153	0.150	0.243
广告点击 Click advertising	16.485	193.765	144.90	34.572
用户违约 User default	0.041	0.010	0.059	0.076
疝气病症 Colic symptoms	0.022	0.004	0.021	0.037
合计 Sum	23.566	474.065	161.261	48.705

3　结论

针对 KNN 方法对样本不均衡数据集的预测偏差问题,本文通过使用 SOMTE 方法对少数样本类的训练集进行采样操作,使得训练集的正负样本数量保持基本一致,从而改善了不均衡数据对 KNN 模型预测性能的影响。对比实验结果发现,基于 SMOTE 改进后的 KSID 方法,比逻辑回归方法、原始 KNN 方法、SVM 决策树方法的分类效果更优,对于较大数据集其时间复杂度更小。但对于样本量和特征维度较大的数据集(如广告点击数据集),KSID 方法使用 SMOTE 采样会产生较多的伪样本,导致影响了模型对原始样本的分类性能,使得 KSID 方法的分类性能低于 SVM 决策树方法。

参考文献

[1] 沈焱萍,伍淳华,罗捷,等. 基于元优化的 KNN 入侵检测模型[J]. 北京工业大学学报,2020,46(1):24-32.

[2] 王轶凡. 考虑数据时效性的高效 KNN 方法[J]. 赤峰学院学报:自然科学版,2019,35(11):19-21.

[3] 王志华,刘绍廷,罗齐. 基于改进 K-modes 聚类的 KNN 分类方法[J]. 计算机工程与设计,2019,40(8):2228-2234.

[4] 张万桢,刘同来,邬满,等. 使用环形过滤器的 K 值自适应 KNN 方法[J]. 计算机工程与应用,2019,55(23):45-52,85.

[5] 余鹰,苗夺谦,刘财辉,等. 基于变精度粗糙集的 KNN 分类改进方法[J]. 模式识别与人工智能,2012,25(4):617-623.

[6] 樊存佳,汪友生,边航. 一种改进的 KNN 文本分类方法[J]. 国外电子测量技术,2015,34(12):39-43.

[7] 罗贤锋,祝胜林,陈泽健,等. 基于 K-Medoids 聚类的改进 KNN 文本分类方法[J]. 计算机工程与设计,2014,35(11):3864-3867,3937.

[8] BLAGUS R, LUSA L. SMOTE for high-dimensional class-imbalanced data [J]. BMC Bioinformatics, 2013, 14(1):106.

[9] 张士翔,李汪根,李童,等. 一种改进的贝叶斯逻辑回归核心集构建方法[J]. 计算机科学,2019,46(S2):98-102.

[10] 刘淑英,邹燕飞,李依桥,等. 基于 KNN 方法的约会网站配对模型的应用研究[J]. 数字技术与应用,2019,37(10):128-129.

[11] 钟龙申,高学军,王振友. 一种新的基于 K-means 改进 SMOTE 方法在不平衡数据集分类中的应用[J]. 数学的实践与认识,2015,45(19):198-206.

[12] 黄勇,魏乐. 一种针对不均衡数据集的 SVM 决策树方法[J]. 成都信息工程大学学报,2019,34(3):274-277.

[13] GOIN J E. Classification bias of the *k*-nearest neighbor algorithm [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(3):379-381.

A SMOTE based KNN Classification Method for Unbalanced Samples

LIN Yongchang, ZHU Xiaoshu
(School of Computer Science and Engineering, Yulin Normal University, Yulin, Guangxi, 537000, China)

Abstract: In order to solve the problem that the prediction result of the KNN method will be biased to the dominant class when the data samples are not balanced, this paper proposes a KNN classification optimization method (KSID) for unbalanced samples based on the synthetic minority oversampling technique (SMOTE). Firstly, this method uses the SMOTE method to equalize the unbalanced training set and the logistic regression model is trained. Secondly, the logistic regression model is used to predict the training set, and the data predicted as positive samples is obtained. The SMOTE method is used to equalize the positive samples and train the KNN model. Finally, the test set is put into the KNN model combined with the logistic regression method for prediction, and the final prediction result is obtained. Based on six unbalanced data sets, KSID is compared with logistic regression, KNN, and SVM decision trees. The results show that the KSID method is superior to the other three methods in the four performance indicators of accuracy, recall, precision, and $F1$ score. By introducing SMOTE, the KSID method overcomes the problem of classification bias when KNN encounters an unbalanced sample data set, and provides a reference for further research on the optimization and application of the KNN method.

Key words: unbalanced sample, KNN, SMOTE, KSID, Logistic regression, classification

责任编辑:符支宏