基于伽马内核与加权 K 近邻的流量分类算法

徐　魁[1],海　洋[1],许艺凡[2,3],段靖海[2],孙炜策[2,3],陶　军[2,3]
(1. 宝鸡市公安局通信处,陕西 宝鸡 721014;
2. 东南大学 网络空间安全学院,江苏 南京 211189;
3. 计算机网络和信息集成教育部重点实验室(东南大学),江苏 南京 211189)

摘　要:K 最近邻算法(KNN)是一种简单有效的分类方式。当数据集分布均衡,不同类别样本之间的差异显著时,KNN 的分类效果一般较好。但实际中数据集通常不理想,网络流量往往呈现倾斜分布,存在样本之间差异不显著等问题。为了更好地权衡样本距离之间差异以及流量类别分布不均带来的模型准确率下降问题,提出了一种基于 Gamma 内核与加权 KNN 的流量分类算法,综合考虑了距离和流量分布对分类结果的影响。采用 Gamma 分布函数作为内核,对不同类别采用自信息进行加权。最后得到 G-WKNN 模型,并将该模型应用于 CIC-IDS2017 数据集。实验结果表明,在流量均衡的情况下,模型准确率稳定在 0.91 左右。在流量不均衡时,依旧具备良好的分类表现。对比其余几种改良的 KNN 算法,其分类准确率较高且模型稳定性好,对 K 值相对不敏感。同时 G-WKNN 模型对少数类别分类准确率的提升效果也较为显著。

关键词:K 最近邻算法;Gamma 分布;自信息;距离函数;网络流量分类
中图分类号:TP393　　**文献标识码**:A　　**文章编号**:1673-629X(2023)02-0214-07
doi:10.3969/j.issn.1673-629X.2023.02.032

Traffic Classification Algorithm Based on Gamma Kernel and Weighted K-Nearest Neighbors

XU Kui[1], HAI Yang[1], XU Yi-fan[2,3], DUAN Jing-hai[2], SUN Wei-ce[2,3], TAO Jun[2,3]
(1. Baoji Municipal Security Bureau, Baoji 721014, China;
2. School of Cyber Sci. & Engr., Southeast University, Nanjing 211189, China;
3. Key Lab of CNII, MOE (Southeast University), Nanjing 211189, China)

Abstract: K-Nearest Neighbors (KNN) is a simple and effective way of classification. When the distribution of the dataset is balanced and the differences between samples of different categories are significant, the classification effect of KNN is generally good. However, the dataset is usually not ideal, and network traffic tends to present skewed distribution, with insignificant differences between samples and other problems. To better balance the difference between sample distances and the problem of model accuracy degradation caused by uneven distribution of traffic categories, we propose a traffic classification algorithm based on Gamma kernel and weighted KNN, which comprehensively considers the impact of distance and traffic distribution. The Gamma distribution function is used as the kernel, and the self-information is weighted for different categories. Finally, G-WKNN model is obtained and applied to the CIC-IDS2017 dataset. The experimental results show that the model accuracy is stable around 0.91 in the case of balanced traffic. When the traffic is unbalanced, it still has a good classification performance. Compared with the other improved KNN algorithms, its classification accuracy is higher with better stability, and more insensitive to the K value. At the same time, the G-WKNN model has a significant improvement effect on the classification accuracy of minority categories.

Key words: KNN; Gamma distribution; self-information; distance function; network traffic classification

0　引　言

网络安全逐渐成为当今信息化社会的焦点问题,目前网络上存在的攻击种类相当多,例如拒绝服务攻击、端口扫描以及 Web 木马等等。异常流量的分类成为防护环节中必不可少的一环。除了传统基于端口和载荷的流量分类,将机器学习应用至网络流量的分

收稿日期:2022-04-20　　**修回日期**:2022-08-20
基金项目:中国高校产学研创新基金-阿里云高校数字化创新专项(2021ALA03006)
作者简介:徐　魁(1973-),男,研究方向为大数据分析;通讯作者:陶　军(1975-),博士,教授,博导,CCF 高级会员(42637S),研究方向为网络安全、物联网技术等。

类[1]也越来越多。相关的机器学习算法包括了聚类算法、KNN 算法、决策树、随机森林和深度神经网络等[2]。其中,在分类问题中,KNN 算法简单而高效。

对于各种分类问题的改进,主要体现在权重函数改良、距离函数的设计等方向;而数据集不均衡问题的处理方式最常见的是过采样和降采样,让数据集分布变得均衡,但通过加权的方式也能够抑制数据集不均衡带来的问题。

在权重设计方面,H. Yigit 等[3]提出用人工蜂群算法找到 KNN 的最佳权重,从而实现更好的分类效果;Su 等人[4]提出使用遗传算法来计算最佳权重,同时用聚类中心代替部分训练数据集,缩短执行时间;毕等人[5]提出了一种基于高斯函数定权的 KNN 算法,应用在室内定位问题上,提高了定位准确率和鲁棒性。

而在距离函数的处理上,Juan Aguilera 等人[6]提出了用基于牛顿引力公式进行加权;Zhang 等人[7]提出了基于随机森林的加权特征,考虑了每个特征的差异性;但两者都没有考虑数据集不均衡问题。杨等人[8]提出了一种基于语义距离的 KNN 算法,分析属性内取值的语义差异,尤其在数据集不完整的情况下,分类准确率优于传统方法。

在函数设计方面,Lin 等人[9]提出了 Focal Loss 的概念,在原有的交叉熵损失函数上做了进一步修改,使得模型更加关注于小样本的部分,进一步提高了网络分类的准确率,这一修改思路也可以在 KNN 分类中借鉴;周等人[10]提出了一种基于聚类改进的 KNN 分类算法,先聚类再分类,提升了 KNN 算法的效率;韦等人[11]提出了一种基于改进极端随机树的方法,通过对加权统计和修改投票规则,其模型在异常流量样本较少的类别上分类效果较好。

对于数据集不均衡问题,Liu 等人[12]提出了类置信权重,依据每类标签属性值的概率对 KNN 进行加权,对于不平衡的数据集,其分类效果有一定的改善;Cao 等人[13]为每个样本分配反比例权重然后与距离权重结合,这是一种简单的距离加权;Ying 等人[14]在异常日志分类中用到了大量数据挖掘和自然语言处理的方法,使数据集不平衡问题得到改善。

Zhao 等人[15]提出了一种加权混合集成方法,在 Boosting 框架下,为每个基分类器分配权重,实现对不平衡数据集的有效分类;Li 等人[16]提出了一种融合的等比例抽样方法,对少数类进行了更多的关注,提升了分类的准确率;王等人[17]提出了一种 SVM 和 KNN 结合的算法,对于较远样本使用 SVM 分类,对于距离近的样本用支持向量进行 KNN 分类,提高了少数类样本的识别率。

1　KNN 算法

在实际网络环境中,流量往往呈现倾斜式分布,异常流量是少数,数据集一般不均衡,这对于分类结果产生了一定的影响。

原始的 KNN 根据预测样本周围距离最近的 K 个点,根据少数服从多数的原则,将这个数据划分为 K 个数据中出现次数最多的那个类别。

若 K 的取值较小,分类器很容易受到局部噪声数据的干扰,分类会出现较大偏差;如果 K 的取值过大,分类器忽略了训练数据中的有效信息,样本更容易被分类到数目多的类别中。而数据集不均衡更容易导致分类准确率下降。

传统的 KNN 分类存在一定的缺陷,例如,没有充分考虑距离远近对样本权重的影响,距离待测样本近的点,其对分类的结果影响应该更大,应当有更大的权值;同时在数据集分布不均时,KNN 总是倾向于将未知测试样本分类为占比数量较多的类,导致占比数量较少的类的分类效果不佳。而在网络流量中,异常流量本身属于少数类别,少数类别的分类准确率对于整个模型分类准确率的贡献也很重要,但传统的 KNN 不能很好地区分出这些少数异常流量的样本。

而简单的 WKNN,为每一个训练样本点给予 d 的权重,其中 d 是待测样本与训练样本的距离。这样的方法一定程度上给近距离的点更大的权重,但也使得计算量变大,同时当 d 趋近于 0 时,即待测样本和训练样本非常接近时,其权重 $1/d$ 会趋近于无穷,对应权重值没有上确界。事实上,在高维数据中,计算得到的距离通常都较小,样本之间的差异性更难体现。若这一点为噪声数据,反而会使得分类准确率下降。同时,简单的 WKNN 也不能缓解数据集分布不均带来的问题。

对于上述提到的两个问题,一个是分类样本对距离的感知所有差别,不同距离的样本,其权重应该要有区分,这是区分样本间的差异性;另一个是对于数据集不均衡的处理,由于 KNN 总是倾向于将预测样本分类为数量较多的类别,因此需要动态调整权重,这是类别间差异性的体现。

同时,KNN 算法属于计算密集型,记训练数据的数量为 n,数据的维度为 m,则预测一个样本,其时间复杂度 $T = O(mn)$。若在这一阶段进行加权,与一个训练样本每计算一次距离,则至少需要多 m 次乘法运算,预测一个样本类别至少多出 mn 次乘法运算。而 KNN 分类的实质是考虑其周边很少的点,这些点称为关键决策点;对于那些距离很远的点,实则参考性并不大,因此在第二个“投票阶段”进行加权和处理。由于 $K \ll n$,因此计算量也会随之减少。

为了解决这些问题,该文提出了一种基于 Gamma

分布函数的自适应加权 KNN 模型(G-WKNN),对于数据集的分布具有一定的自适应性,能够根据样本的差异和类别的差异综合对分类模型进行调整,在提高模型准确率和稳定性的同时,对数量少的类别的分类效果也有提升。在数据不均衡时,类别负载因子起主要调节作用;而在数据较为均衡时,距离感知因子起主要调节作用,从而动态适应数据的变化。

2 基于 Gamma 分布的自适应 WKNN

针对上述提到的问题,该文提出了基于 Gamma 内核的自适应加权 KNN 算法,即 Gamma-WKNN(G-WKNN)。其中,综合权重 w 由距离感知权重 w_d 和类别负载权重 w_f 构成。距离感知权重 w_d 反映了决策样本的差异,能够起到不同样本间的调节作用;类别负载权重 w_f 反映了各个类别之间的差异,对于类别不均衡问题,能够起到调节作用。综合权重 w 的计算方法如式(1)。

$$w = w_d w_f \tag{1}$$

2.1 距离感知权重 w_d

记数据集一共有 C 个类别,样本具有 M 个属性,在空间 R^m 中,测试样本 $P=(x_1,x_2,\cdots,x_m)$ 和训练样本 $Q=(y_1,y_2,\cdots,y_m)$ 间的闵可夫斯基距离 dist 可以表示为:

$$\text{dist}=\left(\sum_{i=1}^{m}|x_i-y_i|^p\right)^{\frac{1}{p}} \tag{2}$$

p 取 1 或者 2 对应的闵式距离最为常用,这里取 $p=2$,此时的距离为欧几里得距离,得式(3):

$$\text{dist}=\|P-Q\|=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \tag{3}$$

距离感知权重 w_d,指标应该呈现出随着距离增大,权重降低。同时所选函数应该有上确界,在 d 很小,甚至趋近于 0 时,不应该出现无穷大的情况,需要避免 $1/d$ 这样的函数的缺陷,其衰减应该控制在一个区间范围内;对于常见的高斯内核而言,其可调参数有 (μ,σ),但由于高斯内核的对称性,只能够使用内核函数的一部分区间,同时高斯内核的参数 μ 只影响曲线在 x 轴上的偏移,因此可调参数仅有 σ。相比之下,Gamma 内核的变化较为丰富,更能够适应不同的情况。Gamma 内核定义如式(4)所示:

$$\text{Gamma}(x\mid\alpha,\beta)=\frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \tag{4}$$

其中,参数 α 决定了曲线的形状;而 β 决定了曲线的陡峭程度,β 越大,曲线越陡峭。相比于反比例函数,Gamma 分布函数随距离衰减程度可控,可以自由调整;相比于高斯内核,其分布不完全对称,具有一定的扰动,这样的随机性一定程度能够避免受到噪声数据的影响,减少过拟合。

对式(4) x 求导,得式(5):

$$\text{Gamma}(x\mid\alpha,\beta)'=\frac{\beta^{\alpha}}{\Gamma(\alpha)}e^{-\beta x}x^{\alpha-2}(\alpha-1-\beta x) \tag{5}$$

令 $\text{Gamma}(x\mid\alpha,\beta)'=0$,得到方程的根,这里考虑 $\alpha\geqslant 1,\beta>0$ 的情况,得式(6):

$$x_0=\frac{\alpha-1}{\beta} \tag{6}$$

结合一阶导数的在 x_0 处的符号和函数图像,可以判定 x_0 为一个极大值点。将 x_0 带入式(4),得到 Gamma 的极大值 $\text{Gamma}(\alpha,\beta)^*$,如式(7)所示:

$$\text{Gamma}(\alpha,\beta)^*=\frac{\beta(\alpha-1)^{\alpha-1}e^{-(\alpha-1)}}{\Gamma(\alpha)} \tag{7}$$

关于参数 α,β 对函数峰值的影响,$\text{Gamma}(\alpha,\beta)^*$ 函数是关于 (α,β) 的单调函数,关于 β 单调递增,关于 α 单调递减。即,若希望函数峰值较大,可以选择较小的 α 和较大的 β。

当 $\alpha>1,\beta>0$ 时,可以得到 Gamma 分布函数导数的递推公式,如式(8)所示:

$$\text{Gamma}(x\mid\alpha,\beta)'=\left(\frac{\alpha-1}{x}-\beta\right)\times\text{Gamma}(x) \tag{8}$$

根据递推关系式(8),可求得二阶导数,如式(9)所示:

$$\text{Gamma}(x\mid\alpha,\beta)''=\left(\frac{\beta x-(\alpha-1)^2-(\alpha-1)}{x^2}\right)G(x) \tag{9}$$

令 $\text{Gamma}(x\mid\alpha,\beta)''=0$,可以求出其变化率最快的点 x_m,如式(10):

$$x_{m\pm}=\frac{\pm\sqrt{\alpha-1}+(\alpha-1)}{\beta} \tag{10}$$

在区间 $[x_0,x_{m+}]$ 内,既能够保证 Gamma 分布函数值较大,同时也能够保证距离衰减速率较快,从而让距离衰减权重真正发挥作用。

综上,距离感知权重 w_d 的计算公式如式(11):

$$w_d=\frac{\text{Gamma}(\text{dist}+x_0\mid\alpha,\beta)}{\text{Gamma}(\alpha,\beta)^*} \tag{11}$$

其中,$\text{dist}\geqslant 0$。

2.2 类别负载权重 w_f

记数据集一共有 N 个样本,每个类别的样本数为 $C_i(i=1:m)$。假设抽样是随机的,则用于训练的样本的分布和总体的分布是一致的,考虑不同类别数量的差异性,假设不同类别的数据之间相对独立,其出现的概率分布 $P=\{p_1,p_2,\cdots,p_n\}$。其中 p_i 的计算方式如式(12)所示:

$$p_i = \frac{C_i}{N} \tag{12}$$

则类别的自信息 $I_e = -\log(p_i)$,即类别被选中的概率越小,其在实际决策的时候,所蕴含的自信息越大,即事件发生时,其能够减少不确定性的程度越大。在考虑类别负载权重时,对于自信息越大的类别应该更加重视,其权值也更大。

令 $w_{fi} = I_e$,得式(13):

$$w_{fi} = \log(\frac{N}{C_i}) \tag{13}$$

对于这样一个决策系统,系统自信息的期望值 H 如式(14)所示:

$$H = E(-\log(p_i)) = -\sum_{i=1}^{m} p_i \log(p_i) \tag{14}$$

实际上, H 是这个系统的熵值,熵值越大,其在决策前,不确定性越大。容易证明,当 $p_i = 1/m$ 时,即每种类别出现的概率一致时, H 有最大值 $H_0 = \log(m)$ 。此时外部数据的噪声对系统影响最小。即,在实际决策时,熵的减少最大限度依赖于 KNN 模型的有效计算,而不是初始的系统分布。特别地,如果数据集分布不均,类别预测时,会较大程度受到系统初始分布的影响,KNN 决策时对熵减小的贡献度有损失。

一般 KNN 的 K 的取值范围为 3 ~ 15,不会过大,过大容易倾向于分类为数量较多的类别;也不宜过小,否则容易受到噪声数据干扰。

函数在前期变化剧烈,当类别比例差距足够大时,变化相对缓和。例如,两个类别的数据量差距在 100 倍左右时,权值差距在 4.5 倍左右;而类别数据量差距在 1 000 倍左右时,权值差距在 7 倍左右。权值差距倍数与超参数 K 的取值相当,能够较好权衡类别数量差距带来的影响,从而实现根据数据集的差异进行动态加权。

2.3　G-WKNN 算法流程

记有 n 个训练样本,样本的维度数为 m ,一共有 M 个类别 $C_1 \sim C_M$,算法输入包括了训练数据 $X_{n,m}$,测试样本 $P_{1,m}$ 。算法实现过程如算法 1 所示。

算法 1:G-WKNN 算法实现

输入:训练样本数据 X ,标签 Y ,以及待测样本 P

输出:待测样本的分类结果 C_i

1. 初始化 K 、α 和 β
2. FOR i to n do
3. $\text{dist}_i \leftarrow 0$
4. FOR j to m do
5. $\text{dist} \leftarrow \| X_i - P \|$ /* 计算 X_i 与 P 之间的欧几里得距离 */
6. END FOR
7. END FOR
8. $\text{dist} \leftarrow \text{Sort}(\text{dist})$ /* 对距离向量升序排序 */
9. 选择 dist 的前 K 个元素
10. 初始化 score ,每个类别均为 0
11. FOR k to K do
12. /* $f(k)$ 是第 k 个元素对应的类别映射 */
13. $\text{score}_{f(k)} \leftarrow \text{score}_{f(k)} + w_{f(k)} \times \text{Gamma}(\text{dist}_k + x_0 \mid \alpha, \beta)/\text{Gamma}(\alpha,\beta)$ *
14. END FOR
15. $\text{score} \leftarrow \text{ArgSort}(\text{score})$ /* 对 score 降序排序,返回下标 */
16. RETURN $f(\text{score}[0])$

3　实验与结果

为了验证 G-WKNN 模型,使用公开数据集对模型的准确率、稳定性以及参数灵敏度等指标进行了对比实验。

3.1　实验数据集

CIC-IDS2017 数据集从 2017 年 7 月 3 日(星期一)开始搜集,于 2017 年 7 月 7 日(星期五)结束。主要攻击类型包括暴力破解 FTP 和 SSH、DoS、Heartbleed、Web 攻击、渗透、僵尸网络和 DDoS。

数据集正常流量类数目相对较多,流量呈现高度分布不均,正常流量占到了 80%,对于部分异常流量,其占比不超过 1%。对数据集做了如下调整,剔除了部分数据量极少的异常流量,如 Sql Injection 类型的异常流量,同时对正常流量抽取一部分作为实验数据。

3.2　评价指标

实验中使用了混淆矩阵、准确率、精确度、召回率和 F1 值对模型进行评估。

其中准确率的定义如式(15):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{15}$$

精确度表示预测为正样本中,被正确分类的比例,如式(16)所示:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

召回率表示对于真实的样本,预测为正样例所占比例,如式(17)所示:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

而 F1 分数兼顾了模型的精确率和召回率,其表达式如式(18)所示:

$$\text{F1}_{\text{score}} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{18}$$

其中,TP、FP、TN 和 FN 的含义如下:

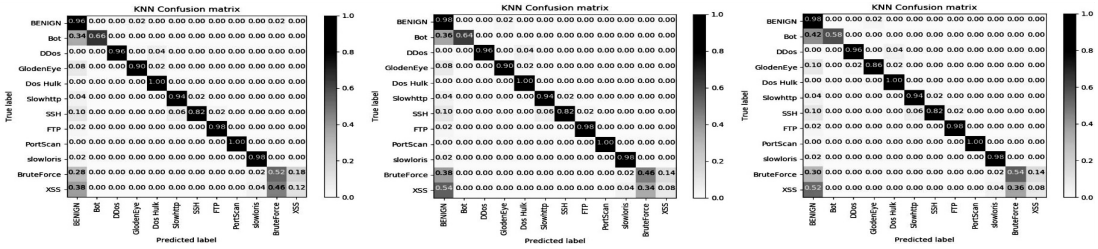
在二分类的情况下,TP 表示正例被预测为正例;FP 表示正样例被错误识别为负样例;TN 表示负样例被正确识别为负样例;而 FN 表示负样例错误地被预测为正样例;多分类计算方式同理。

3.3 少数类分类性能测试

从实验数据集中按类别比例抽取 10 000 条数据作为实验数据。其中数据集不均衡对分类准确率的影响主要体现在训练数据上,即那些被标记好用于未知样本分类的数据,实验中随机按类别比例抽取数据集的一部分作为训练数据,剩下的数据再按照分层抽样,每个类别抽取固定数量的数据用于结果测试。

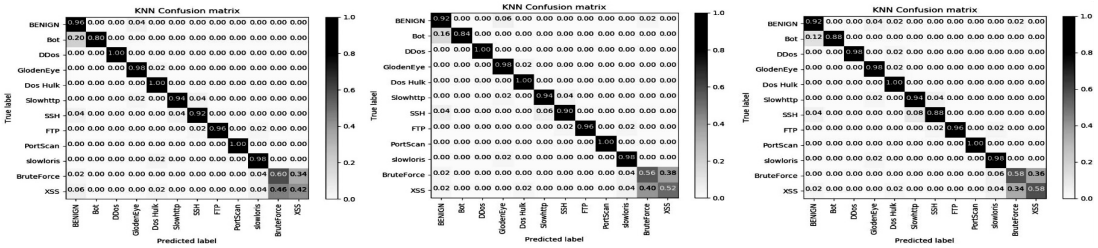
使用原始的 KNN,在 K 取值为 3、5 和 7 的情况下,测试结果的混淆矩阵如图 1(a)、(b)、(c)所示。在数量较少的类别,如 Bot、BruteForce 和 XSS 三类攻击的分类效果相对较差,分别对应于混淆矩阵的第二行和最后两行。尤其是 XSS 攻击类别,对应的混淆矩阵的值仅有 0.04 左右。

在固定参数 $\alpha = 4$、$\beta = 5$ 的情况下,G-WKNN 的表现如图 2(a)、(b)、(c)所示。除了总体分类效果有提升外,对于数量较少的类别,其分类准确率有比较明显的提升。Bot、BruteForce 和 XSS 三类攻击对应混淆矩阵的值达到了 0.8、0.6 和 0.5 左右。混淆矩阵的对角线更加突出,模型对少数类别分类的准确率有提升。



(a) $K = 3$ (b) $K = 5$ (c) $K = 7$

图 1 原始 KNN 分类混淆矩阵



(a) $K = 3$ (b) $K = 5$ (c) $K = 7$

图 2 G-WKNN 分类混淆矩阵

3.4 参数 α 和 β 灵敏度分析

为了对比不同 α 和 β 参数对模型分类效果的影响,本节设计了参数 α 和 β 敏感性分析实验,以验证模型的稳定性和解释参数选择的合理性。

选择不同 α 和 β,测试模型的分类准确率,对比不同 K 值情况下模型的稳定性,实验结果如图 3 所示。

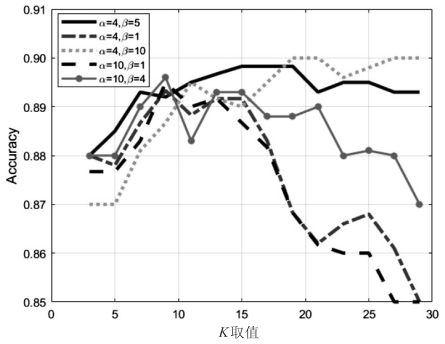


图 3 α 和 β 参数对模型准确率影响

参数 K 在 9 至 17 时,参数 α 和 β 的选择对模型准确率影响不显著,模型准确率基本维持在 0.89 上下;但在 K 值较小或者较大时,α 和 β 对模型分类准确率影响变大。

实验数据计算出的距离值分布在[0,1]内,故总体而言,衰减区间 $[x_0, x_{m+}]$ 的长度 $L = \sqrt{\alpha - 1}/\beta$ 在 0 至 1 之间,并且相对较小时,模型相对稳定,准确率较高。因此,$\alpha = 4$、$\beta = 5$ 和 $\alpha = 4$、$\beta = 10$ 两组对应的准确率较高,模型相对更加稳定。

同时,无论 α 和 β 的取值如何变化,模型准确率较高的点首次出现于 $K = 9$ 前后,虽然一些模型随着 K 增大,其准确率还会略有增长,但不明显。当 K 值过大时,此时权值差距倍数与 K 值相差较大,模型的准确率会受到一定程度影响;尤其对于衰减区间长度较长的模型,其准确率下降较为显著,此时模型动态加权的能力减弱。

3.5 参数 K 灵敏度分析

K 是 KNN 模型关键的超参数。为此,本节设计了不同模型对参数 K 灵敏度分析实验,以验证 G-WKNN 模型的有效性。

固定参数 α 和 β,对比分析了几种改良 KNN 模型的灵敏度,测试数据如表 1 所示。总体而言,加权 KNN 的表现比原始 KNN 更好;而 G-WKNN 的表现更佳,比 $1/d$ 型的 KNN 准确率高,并且受超参数 K 的影响更小,同时对于超参数 α 和 β,模型表现不过于

灵敏。

表 1　实验结果

类型	K 值	F1	Recall	Accuracy
原始 KNN	3	0.80	0.81	0.80
	5	0.78	0.78	0.775
	7	0.77	0.77	0.768
原始 $1/d$ 型 WKNN	3	0.81	0.82	0.82
	5	0.81	0.81	0.81
	7	0.81	0.81	0.808
$1/d$ 替换 Gamma-WKNN	3	0.85	0.85	0.853
	5	0.85	0.85	0.853
	7	0.85	0.85	0.853
$\alpha = 4, \beta = 5$ G-WKNN	3	0.88	0.88	0.88
	5	0.88	0.89	0.885
	7	0.88	0.89	0.893
$\alpha = 4, \beta = 10$ G-WKNN	3	0.87	0.88	0.877
	5	0.88	0.88	0.876
	7	0.89	0.89	0.888

就模型分类准确率而言,对 K 值从 3 到 21 逐一测试,原始 KNN 和三种 WKNN 的对比结果如图 4 所示。

G-WKNN 的准确率基本维持在 0.9,对 K 选择敏感度低;其余两种 WKNN 相对稳定,但准确率在 0.85 和 0.8 左右;而原始 KNN 对 K 的选择较为敏感,模型预测准确率容易随着 K 的取值发生较大波动;G-WKNN 具备较好的调节能力,其准确率波动更小,模型更加健壮。

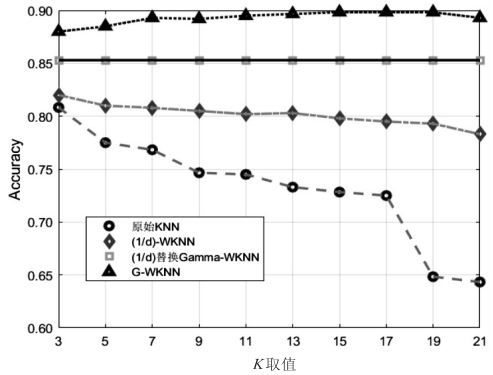


图 4　K 取值对预测准确率的影响

将 $1/d$ 加权方式替换 Gamma 内核,替换后的模型,其分类准确率比原始的 $1/d$ -WKNN 更好,更稳定,证明了类别负载权重调节的有效性;但准确率比 G-WKNN 低,说明了在合适的参数 α 和 β 下,Gamma 分布函数对于样本距离加权调节更加有效。总体而言,G-WKNN 分类准确率的波动更小,其准确率比另外两种 WKNN 也更高。

3.6　数据集均衡分布对模型的影响

为了说明不均衡数据集对模型准确率的影响,本节设计了分层抽样实验,通过分层抽样,能够将每个类别的训练数据集控制在相当的范围,测试样本不变。由于 XSS 类别数目最少,仅有 600 余条,选择 75% 作为训练集,各个类别采样均为 450 条。此时,不同 KNN 模型分类结果如图 5 所示。

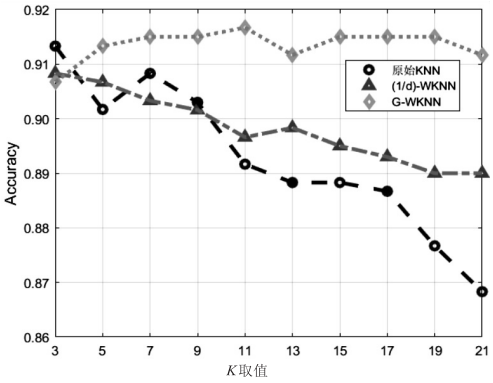


图 5　数据集分布均衡下各个模型分类准确率

从图 5 可以看出,在数据集分布均衡的情况下,三种模型的分类准确率均有不少提升,但从稳定性和分类结果看,原始 KNN 和 $1/d$ -WKNN 对于超参数 K 较为敏感,模型分类准确率不稳定,而 G-WKNN 分类准确率基本维持在 0.91 以上。

在类别均衡的情况下,类别负载权重 w_f 作用被削减;特别地,若类别绝对均衡,则类别负载权重退化为 1,此时距离感知权重发挥调节作用,从实验结果可以看到,Gamma 内核对分类器的调节是有效的,G-WKNN 分类准确率始终高于其他两个模型。

3.7　模型运行时间性能分析

为了说明模型运行时间性能,本节设计了对比实验。实验数据集数量为 10 000 条,在测试数据集为每个类别 50 例的情况下,对比了原始 KNN 和 G-WKNN 的运行时间,如表 2 所示。

表 2　模型运行时间实验结果

运行时间/s	$K = 3$	$K = 5$	$K = 7$	$K = 9$
原始 KNN	43.7	48.4	47.2	48.7
G-WKNN	46.9	47.4	45.8	45.7

由表 2 可知,G-WKNN 在运行时间上尽管优势不显著,与原始 KNN 运行时间相当,但模型在准确率上的提升较为突出。

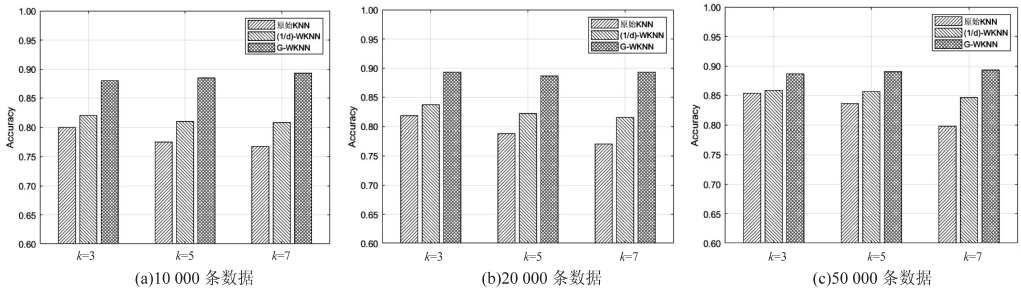
3.8　数据集数量对模型的影响

为了说明数据量大小对模型的影响,本节设计了对比实验,通过调节参与分类的数据集的数量,对比不同模型的分类准确率。

实验结果表明,通过增加训练数据集,一定程度上缓解了数据集不均带来的问题,如图 6(a)、(b)和(c)所示。增加训练数据集后,三种模型分类准确率有一定的提升,原始的 KNN 通过增加数据集准确率最高

达到了 0.85;而 1/d -WKNN 随训练样本数量的增加,分类效果也有一定的提升;G-WKNN 模型的准确率基本维持在 0.88 至 0.90。这说明,通过增加数据集数量的方式,一定程度能够缓解数据集不均带来的分类准确率下降的问题。

但在实际情况中,数据集往往分布不理想,同时数据集数量可能有限,G-WKNN 算法能够在数据集较少,并且分布不平衡的情况下,达到比较高的分类准确率。数据集数量对模型准确率影响较小,模型较为稳定。



(a)10 000 条数据 (b)20 000 条数据 (c)50 000 条数据

图 6 三种模型在不同数据量下的分类准确率对比

4 结束语

G-WKNN 算法缓解了数据集分布不均带来的类别倾向问题。通过设计距离感知权重和类别负载权重,分别调节类别间的差异和样本间的差异。对于类别分布不均的数据集能够自动适应数据集的倾斜程度,其表现优于传统 KNN 和简单的 1/d -WKNN,并且结果更加稳定;在分类结果中,针对数目较少的类别,其预测的准确率有比较明显的提升。这对于倾斜分布的数据集以及数据集较少的情况,分类效果有较为明显的改善。同时,分类效果也不能仅仅依赖于权重函数的设计,实验结果也表明,在相同的情况下,分布均匀的数据集,分类器的表现会更加突出和稳定。

参考文献:

[1] 张 蕾,崔 勇,刘 静,等. 机器学习在网络空间安全研究中的应用[J]. 计算机学报,2018,41(9):1943-1975.

[2] NGUYEN T T T,ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. IEEE Communications Surveys & Tutorials,2008,10(4):56-76.

[3] YIGIT H. A weighting approach for KNN classifier[C]// 2013 international conference on electronics, computer and computation (ICECCO). Ankara:IEEE,2013:228-231.

[4] SU M Y. Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification[J]. Journal of Network and Computer Applications, 2011, 34(2):722-730.

[5] 毕京学,甄 杰,汪云甲,等. 高斯函数定权的改进 KNN 室内定位方法[J]. 测绘通报,2017(6):9-12.

[6] AGUILERA J, GONZÁLEZ L C, MONTES-Y-GÓMEZ M, et al. A new weighted k-nearest neighbor algorithm based on newton's gravitational force[C]//Iberoamerican congress on pattern recognition. Madrid: Springer, 2018: 305-313.

[7] ZHANG H, BU F. Weighted KNN algorithm based on random forests[C]//Proceedings of the 2019 11th international conference on machine learning and computing. Zhuhai:[s. n.],2019:231-235.

[8] 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.

[9] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017: 2980-2988.

[10] 周庆平,谭长庚,王宏君,等. 基于聚类改进的 KNN 文本分类算法[J]. 计算机应用研究,2016,33(11):3374-3377.

[11] 韦海宇,王 勇,柯文龙,等. 基于改进极端随机树的异常网络流量分类[J]. 计算机工程,2018,44(11):33-39.

[12] LIU W,CHAWLA S. Class confidence weighted KNN algorithms for imbalanced data sets[C]//Pacific-Asia conference on knowledge discovery and data mining. Shenzhen: Springer,2011:345-356.

[13] CAO Qimin, LA Lei, LIU Hongxia, et al. Mixed weighted KNN for imbalanced datasets[J]. Int J Performability Eng, 2018,14(7):1391-1400.

[14] YING S, WANG B, WANG L, et al. An improved KNN-based efficient log anomaly detection method with automatically labeled samples[J]. ACM Transactions on Knowledge Discovery from Data,2021,15(3):1-22.

[15] ZHAO J,JIN J,CHEN S,et al. A weighted hybrid ensemble method for classifying imbalanced data[J]. Knowledge-Based Systems,2020,203:106087.

[16] LI J,WANG P,LIU X. Research based on unbalanced data set classification algorithm[C]//AIP publishing LLC. Melville:[s. n.],2018:040054.

[17] 王超学,张 涛,马春森. 改进 SVM-KNN 的不平衡数据分类[J]. 计算机工程与应用,2016,52(4):51-55.