doi: 10.12068/j.issn.1005-3026.2019.07.004

面向不平衡数据集的一种改进的 k-近邻分类器

刘 鹏[1,2], 杜佳芝[3], 吕伟刚[2,4], 窦明武[1]
(1. 中国海洋大学 计算中心, 山东 青岛 266100; 2. 中国海洋大学 信息学院, 山东 青岛 266100;
3. 哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001; 4. 中国海洋大学 教育技术系, 山东 青岛 266100)

摘 要: 心脏心律失常数据集的心电图(ECG)数据往往存在各心律失常类型下样本数量不平衡问题. 针对此问题,提出了一种新的模式识别分类方法,即改进的基于核的差重建的加权 k-近邻分类器(modified kernel difference-weighted k-nearest neighbor classifier, MKDF-WKNN),通过引入修正因子对含样本数较多的类别进行权值抑制,对含样本数较少的类别进行权值的加大,并使用 UCI 心脏心律失常数据集对 ECG 数据进行分类. 实验结果表明,提出的算法和其他一些基于 KNN 的算法如 KNN,DS-WKNN,DF-WKNN 和 KDF-WKNN 相比,对于不平衡的心律失常数据集的分类有更好的效果.

关 键 词: 心律失常;心电图;模式分类;k-近邻算法;不平衡数据集
中图分类号: TP 181 **文献标志码**: A **文章编号**: 1005-3026(2019)07-0932-05

A Modified KNN Classifier for Unbalanced Dataset

LIU Peng[1,2], *DU Jia-zhi*[3], *LYU Wei-gang*[2,4], *DOU Ming-wu*[1]
(1. Computing Center, Ocean University of China, Qingdao 266100, China; 2. School of Information, Ocean University of China, Qingdao 266100, China; 3. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; 4. Department of Educational Technology, Ocean University of China, Qingdao 266100, China. Corresponding author: DOU Ming-wu, E-mail: doumingwu@sina.com)

Abstract: The existing arrhythmia datasets are suffering from the unbalanced number of training sample for electrocardiogram(ECG) data due to the obvious difference among the sample number of different types. A novel KNN-based classification algorithm, i.e., a modified kernel difference-weighted KNN classifier(MKDF-WKNN) was proposed, by introducing a correction factor to restrain the weights of the categories with more samples and increase the weights of the categories with fewer samples. The experiment was carried on the UCI arrhythmia dataset to classify the ECG data. The results show that, for unbalanced datasets the proposed algorithm is better than some other KNN-based algorithms such as KNN, DS-WKNN, DF-WKNN and KDF-WKNN, in terms of classification accuracy.

Key words: cardiac arrhythmias; electrocardiogram; pattern classification; KNN algorithm; unbalanced dataset

拥有健康的心脏,是拥有健康身体的必然前提. 近年来,随着人类生活节奏加快,心脏类疾病的发病率逐年提高,成为严重威胁人类生命的疾病之一. 随着计算机技术的突飞猛进,针对心脏病的数字化研究已经逐渐发展起来[1-3]. 在正常情况下,心跳的节律是规律整齐的,但是如果心脏跳动不规律,就称之为心律失常. 心律失常可以被分为若干种类,其中一些心律失常类型是极其危险的,往往暗示着一些潜在的严重的心脏疾病,如果不及时治疗,后果十分严重,甚至会导致突发性死亡[4]. 因此,确诊心律失常的类型并尽早选择有针对性的治疗对于预防和监测心脏病、提高医生的工作效率具有重要意义[5-6].

临床上对心脏类疾病的诊断通常依靠心电图(electrocardiogram,ECG)来完成,心电图是一种间接测量并且容易记录的工具,反映了心脏兴奋

收稿日期: 2018-07-13
基金项目: 山东省自然科学基金资助项目(ZR2017MF051); 教育部人文社科基金资助项目(18YJCZH103).
作者简介: 刘 鹏(1981-),男,山东青岛人,中国海洋大学讲师,博士研究生.

的发生、发展,以及恢复过程,是描述心脏电活动规律的客观指标[7],因此心电图检查对于心律失常的诊断具有重要意义. 在常规的 12 导联心电图中可以获得的有效信息有:① P 波形态和时限是否正常;② QRS 波群的形态和时限;③ PP 间期和 RR 间期的速率和节律;④ P 波与 QRS 波群之间的关系等. 例如,P 波形状不正常,提示冲动的起源不在窦房结;如果 QRS 波群增宽,则需注意其形状是否具有左、右束支传导阻滞的特征;如果 P 波与 QRS 波群无固定关系,则表明有房室分离[8]. 由于心律失常的类型多种多样,因此其对应的心电特征也非常繁杂. 如何根据不同波形的心电图对心律失常进行诊断,进而对症治疗是预防和治疗心律失常的关键[9].

医师通过心电图波形的变化来察觉出心律的不规则变化,从而判定其所属心律失常类型并进一步给出相应治疗方案. 然而,受个人经验不同、病人数量多等因素影响,这种主观的人工诊断方式不仅浪费时间而且容易出错. 因此在心电图诊断的客观化研究中,心电信号模式的自动分类是实现心电诊断自动化的关键. 迄今为止,一系列方法例如心电图信号处理[10]、模式识别[11]、机器学习[12]已经被提出. 同时,一些通用的心脏心律失常数据集在诊断心律失常中也起着重要的作用. 这些心脏心律失常数据可以大致分成两种:信号类型[13]和数值类型[14]. 本文基于数值型的 UCI 心律失常数据集(通过图像识别和信号处理等技术生成的由多维矩阵组成的数据集)来进行分类的研究.

然而,已有的心律失常数据集大都存在不同心律失常类型下样本数量的不平衡问题,因此,直接使用传统的机器学习分类器例如 k – 近邻分类器(k – nearest neighbor classifier, KNN)等进行分类会造成一些较大的误差. 针对这一问题,本文提出了一种新的模式识别分类方法——改进的基于核的差重建的加权 k – 近邻分类器(modified kernel difference – weighted k-nearest neighbor classifier, MKDF – WKNN),并使用 UCI 心律失常数据集进行了精度实验测试和对比实验.

1　相关研究与问题描述

1.1　UCI 心律失常数据集

UCI 机器学习数据集[14]是一个常用的标准机器学习数据库的集合,目前已经被广泛用于机器学习算法的经验分析中. 该数据集目前一共包含了 211 个子数据集,覆盖了生命科学、物理科学、计算机科学工程、社会科学等复杂的研究领域. 本文针对 UCI 数据集中的基于 ECG 的心脏心律失常数据集进行分类研究.

UCI 数据集中心脏心律失常数据集被分成 16 类,分别对应正常情况、心肌缺血、心肌梗死、窦性心动过速、窦性心动过缓等情况,详见表 1,共计 452 个样本. 每一个样本包含 279 个属性,这些属性中,前 4 项是被采集者的一般信息,分别对应年龄、性别、身高、体重,其余的 275 项属性是从标准的 12 导联心电图记录中提取出来的,包括 QRS 波群持续时间等信息. 这里有两点需要注意:第一,由于某些心律失常的发病几率非常低,目前世界上尚无样本数非常均衡的心律失常数据集可供使用,因此,UCI 心脏心律失常数据集中样本对应的类的标签是不平衡的,例如,正常类中有 245 个样本,而在第一度、第二度、第三度房室传导阻滞类中,样本数量都为 0;第二,属性值有缺失,其中缺失的属性信息(约占 0.33%)由“?”替代.

表 1　UCI 心脏心律失常数据集中类的分布

Table 1　Class distribution in UCI cardiac arrhythmia database

类号	对应症状	样本数量
1	正常	245
2	心肌缺血(冠状动脉疾病)	44
3	陈旧性前壁心肌梗死	15
4	陈旧性下壁心肌梗死	15
5	窦性心动过速	13
6	窦性心动过缓	25
7	室性早搏	3
8	室上性早搏	2
9	左束支传导阻滞	9
10	右束支传导阻滞	50
11	房室传导阻滞(第一度)	0
12	房室传导阻滞(第二度)	0
13	房室传导阻滞(第三度)	0
14	左心室肥大	4
15	心房颤动或扑动	5
16	其他	22

1.2　基于 UCI 心律失常数据集的分类器

在对心律失常数据类型分类时,许多例如支持向量机、人工神经网络、k – 近邻算法等经典算法被使用. 近年来还有许多学者也提出了一些基于这些经典算法的改进算法来提高心律失常数据集分类的精度. 其中,k – 近邻分类器(KNN)作为

一种简单、有效、非参数的分类算法,是心律失常分类中较为常用的分类器. KNN 由 Cover 提出,是一个理论上比较成熟的方法[15]. 该算法的基本思想是:根据传统的向量空间模型,数据被形式化为特征空间中的加权特征向量,对于一个测试样本,计算它与训练样本集中每个文本的相似度,找出 k 个最相似的文本,根据加权距离和判断测试文本所属的类别. 基于 KNN 算法很多学者提出了各种改进算法,如基于距离加权 k – 近邻分类器(distance – weighted k-nearest neighbor classifier, DS – WKNN),其通过距离的大小来对 k 个邻居进行赋值并取得了比较好的效果[16]. 文献[17]也提出了一种基于核的差重建的加权 k – 近邻分类器(kernel difference – weighted k-nearest neighbor classifier, KDF – WKNN),并且在 UCI 心律失常数据集上的实验结果也表明此分类器分类精度要优于 DS – WKNN.

1.3 数据样本的不平衡问题

KNN 方法以及其改进算法是一种简单有效的非参数方法,并不需要产生额外的数据来描述规则,它的规则本身就是数据. 但是 KNN 的决策模式决定了该算法受到训练样本的分布状况影响较大. UCI 心律失常数据集,其大部分的样本,即其中的 245 个样本都属于第一类正常类型. 一般说来,不同类别的样本,类内样本数较多类别的样本在 KNN 选择时容易被选中. 为了更好地说明这个问题,这里通过一个在二维空间的例子来展示样本数量不均衡对分类结果的影响,见图 1.

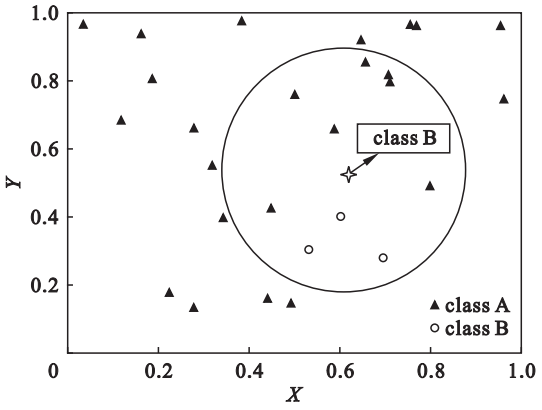

图 1 数据样本不平衡对分类结果影响示意图

Fig. 1 Impact of unbalanced data samples on classification result

如图 1 所示,当使用经典 KNN 算法并取参数 $k=10$ 进行分类时,样本 x_i 就会被分到 A 类,因为该样本的最近的 10 个邻居中 7 个都是 A 类,按照 KNN 决策规则,该样本应该被分为 A 类. 然而,事实上如果该样本属于 B 类,则此分类就是一个由样本数量不均衡导致的错误分类. 并且当 k 取值越大时,这种误差则越大. 为了克服图 1 所展现的问题,本文考虑引入权重系数对大数量类别样本的权重进行抑制,同时加大小数量类别样本的权重. 因此,基于之前提出的 KDF – WKNN 算法,本文进一步提出了一种改进的基于核的差重建的加权 k – 近邻分类器(modified kernel difference – weighted k-nearest neighbor classifier, MKDF – WKNN).

2 改进的基于核的差重建的加权 k – 近邻分类器

给定训练集 $\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_i,\boldsymbol{y}_i),\cdots,(\boldsymbol{x}_m,\boldsymbol{y}_m)\}$,其中 $\boldsymbol{x}_i$ 表示第 i 个训练样本,$\boldsymbol{y}_i$ 表示第 i 个训练样本 $\boldsymbol{x}_i$ 对应的类别标签. 对于一个未知类别样本 $\boldsymbol{x}$,可以用欧几里德距离矩阵来获得样本 $\boldsymbol{x}$ 最近的 k 个邻居 $\{\boldsymbol{x}_1^{\mathrm{NN}},\cdots,\boldsymbol{x}_k^{\mathrm{NN}}\}$. 对于传统的 KNN 决策模式,未知样本的标签可以由这 k 个邻居中类别出现次数最多类别的标签决定. 但是,由于距离这个未知样本较近的那些邻居应该在此样本的分类中起着较为重要的作用,一种基于距离的 DS – WKNN 算法被提出,为每一个邻居依据下面的距离方程赋予一个权值 w_i:

$$w_i=\frac{d(\boldsymbol{x}_1^{\mathrm{NN}},\boldsymbol{x}_k^{\mathrm{NN}})-d(\boldsymbol{x}^{\mathrm{NN}},\boldsymbol{x}_i^{\mathrm{NN}})}{d(\boldsymbol{x}^{\mathrm{NN}},\boldsymbol{x}_k^{\mathrm{NN}})-d(\boldsymbol{x}^{\mathrm{NN}},\boldsymbol{x}_1^{\mathrm{NN}})}. \tag{1}$$

在之前的工作中,k 个邻居的关联度也被纳入分类中并提出了一种差重建的 DF – WKNN 以及基于核的算法,即 KDF – WKNN[17]. 在 DF – WKNN中,对于距离点 $\boldsymbol{x}$ 最近的 k 个邻居 $\boldsymbol{X}=\{\boldsymbol{x}_1^{\mathrm{NN}},\cdots,\boldsymbol{x}_k^{\mathrm{NN}}\}$,其最优权值 w_i 可以通过式(2)来赋权值:

$$\left.\begin{aligned}&\boldsymbol{w}=\arg\min\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\|^2,\\&\text{s.t.}\sum\nolimits_i^{w}w_i=1.\end{aligned}\right\} \tag{2}$$

令 $\boldsymbol{D}=[\boldsymbol{x}-\boldsymbol{x}_1^{\mathrm{NN}},\cdots,\boldsymbol{x}-\boldsymbol{x}_k^{\mathrm{NN}}]^{\mathrm{T}}$,则式(2)中优化问题可表示为

$$\left.\begin{aligned}&\boldsymbol{w}=\arg\min\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{w},\\&\text{s.t.}\sum\nolimits_i^{w}w_i=1.\end{aligned}\right\} \tag{3}$$

令 Gram 矩阵 $\boldsymbol{G}^k=\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$,经过一系列简单的拉格朗日乘数法等数学运算,式(3)可以被化简为

$$\left[\boldsymbol{G}^k=\boldsymbol{G}^k+\frac{\eta\mathrm{tr}(\boldsymbol{G}^k)}{k}\right]\boldsymbol{w}=1_k. \tag{4}$$

其中：$\mathrm{tr}(\boldsymbol{G}^k)$ 是矩阵 $\boldsymbol{G}$ 的迹；正则化系数 $\eta = 1 \sim 1 \times 10^{-3}$. 最终权重系数可以由通过求解方程(3)得到，从而对未知样本进行分类. 对于核空间的 KDF – WKNN，其求解算法也类似于 DF – WKNN，详细内容见文献[17].

本文为了减少样本数量不平衡对 UCI 心律失常数据分类的影响，引入修正因子 $\boldsymbol{c}$，对含样本数较多的类别赋予一个较低的权重对其重要性进行抑制，而对于含样本数较少的类别赋予一个较高的权重. 基于这样的思想，本文修正了 KDF – WKNN 算法得到的权值 $\boldsymbol{w}$. 通过定义 f 为计算某一类内的样本数量的函数，则修正因子 $\boldsymbol{c}$ 可以被归纳为

$$\boldsymbol{c}_i = \frac{\lg\left(\alpha + \dfrac{n}{f(i) + \xi}\right)}{\lg(\alpha + 1)}, \quad i = 1, \cdots, p. \tag{5}$$

这里：n 表示训练集样本数；ξ 是一个可以根据数据样本进行调整的常数参数；$\alpha = \mathrm{round}\ \dfrac{\max(f)}{\mathrm{avg}(f)}$，其中分子分母分别代表所有类的最大样本数和平均样本数. 则最终的权值可被修正为 $\boldsymbol{w}_i := \boldsymbol{w}_i \times \boldsymbol{c}_i$. 从式(5)中可以明显看出，某类数据样本数越大，则修正因子 $\boldsymbol{c}_i$ 越小，那么一个较小的权值会被赋给对应的样本. 当此算法用于 UCI 心律失常数据分类时，第一类“正常”样本由于样本数量较大会被分配一个较小的权值，从而减少其对附近的测试样本的影响. 那么，图 1 中存在的数据样本不平衡对分类结果的误判可以在一定程度上得到控制.

3　实验结果与分析

为了验证本文提出的 MKDF – WKNN 算法对于样本数量不平衡数据分类的有效性，本文与其他一些基于 KNN 的主流分类算法进行了对比实验. 所有的实验都是在处理器为 Intel(R) Core(TM) i5 – 4590(3.30 GHz)的台式机上进行的，编程环境为 Matlab2012a. 数据集选取的是前面介绍的数值型的 UCI 心脏心律失常机器学习数据集. 在心电信息采集和特征提取过程中，由于心电信号特征提取方法失效等原因，数据缺失情况往往不可避免. 为了处理此问题，这里采用处理心律失常数据缺失文献中常用的方法，即用缺失值所在列的平均值来替换此缺失值[18].

在进行分类实验之前，应确定超参数的取值. 对于本实验中所应用的最优超参数 k 近邻的数量 k 和正则化参数 η，采用 10 重交叉验证方法来选取.

这里，UCI 心脏心律失常数据集被随机分成 10 份，然后同样使用 10 重交叉验证方法[19]来计算分类精度. 同时，为了降低性能评估的误差，本研究计算了 10 次 10 重交叉验证方法分类精度的平均值. 所谓 10 重交叉验证方法，具体过程如下：

1）对于给定的一个定义好的训练集 T 以及测试集 S，将训练集 T 分成 10 等份 $\{T_1, T_2, \cdots, T_{10}\}$，其中的每份 T_i 被看作一个新的测试集，同时将其余的 9 份看作训练集；

2）利用分类器在这 10 个子数据集上的平均错误率，找到最优的参数，即平均错误率最小的那组参数；

3）在最优参数确定后，使用训练集 T 重新训练分类器的结构；

4）计算测试集 S 在该数据集上分类器的分类精度.

为了有一个比较全面的评测，这里将 MKDF – WKNN的分类结果与其他一些主流分类方法进行了比较，表 2 列出了 5 种方法的各自的平均分类精度，包括 k – 近邻分类器(KNN)、基于距离的加权 k – 近邻分类器(DS – WKNN)、差重建的加权 k – 近邻分类器(DF – WKNN)、基于核的差重建的加权 k – 近邻分类器(KDF – WKNN)和本文所提出的 MKDF – WKNN 分类器.

表 2　5 种方法的分类准确率

Table 2　Classification accuracy of five methods　　%

KNN	DS – WKNN	DF – WKNN	KDF – WKNN	MKDF – WKNN
58.19	61.50	71.90	71.68	73.01

从表 2 中可以看出，MKDF – WKNN 的分类准确率为 73.01%，比其余 4 种分类方法的准确率都要高. 由此可以看出，本文所提出的算法在对此类样本数量不平衡的 UCI 心律失常数据集分类中可以取得更好的效果.

4 结 论

本文提出了一种改进的基于核的差重建的加权 k – 近邻分类器,即 MKDF – WKNN,用来针对不平衡样本数量心律失常数据集诊断心律失常症状. 对于样本数量不平衡问题,通过引入修正因子对含样本数较多的类别进行权值的抑制,对含样本数较少的类别进行权值的加大,从而在一定程度上处理了不平衡样本数量问题对分类带来的误判. 在性能分析方面使用 UCI 心律失常数据集进行测试. 实验结果表明,相比较 KNN, DS – WKNN, DF – WKNN 和 KDF – WKNN,本文提出的算法有着更高的分类精度.

参考文献:

[1] Bai J, Xie S, Wang K, et al. Simulation research on early after depolarizations-mediated ventricular fibrillation based on a heart model [J]. *Progress in Biochemistry & Biophysics*, 2015, 42(10): 955 – 961.

[2] Yang F, Zhang L, Lu W, et al. Depth attenuation degree based visualization for cardiac ischemic electrophysiological feature exploration [J]. *BioMedical Research International*, 2016, 2016(2): 1 – 8.

[3] Lu W, Li J, Yang F, et al. Effects of acute global ischemia on re-entrant arrhythmogenesis: a simulation study [J]. *Journal of Biological Systems*, 2015, 23(2): 213 – 230.

[4] Jadhav S M, Nalbalwar S L, Ghatol A A. Artificial neural network models based cardiac arrhythmia disease diagnosis from ECG signal data [J]. *International Journal of Computer Applications*, 2012, 44(15): 8 – 13.

[5] Bai J, Wang K, Zhang H. Potential pathogenesis discovery of arrhythmia based on cardiac electrophysiological models: research progress [J]. *Progress in Biochemistry & Biophysics*, 2016, 43(2): 128 – 140.

[6] Lu W, Li J, Yang F, et al. Simulation study of ventricular arrhythmia at the early stage of global ischemic condition [J]. *Progress in Biochemistry and Biophysics*, 2015, 42(2): 189 – 194.

[7] Kumar R G, Kumaraswamy Y S. A neural network approach for cardiac arrhythmia classification [J]. *IUP Journal of Computer Sciences*, 2013, 7(1): 62 – 70.

[8] Jambukia S H, Dabhi V K, Prajapati H B. Classification of ECG signals using machine learning techniques: a survey [C]// Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances. Ghaziabad, 2015: 714 – 721.

[9] Aslanidi O V, Clayton R H, Lambert J L, et al. Dynamical and cellular electrophysiological mechanisms of ECG changes during ischaemia [J]. *Journal of Theoretical Biology*, 2005, 237(4): 369 – 381.

[10] Thakor N V, Zhu Y S. Applications of adaptive filtering to ECG analysis: noise cancellation and arrhythmia detection [J]. *IEEE Transactions on Biomedical Engineering*, 1991, 38(8): 785 – 794.

[11] Coast D A, Stern R M, Cano G G, et al. An approach to cardiac arrhythmia analysis using hidden Markov models [J]. *IEEE Transactions on Biomedical Engineering*, 1990, 37(9): 826 – 836.

[12] Sultana N, Kamatham Y. MSVM-based classifier for cardiac arrhythmia detection [C]// Advances in Computing, Communications and Informatics (ICACCI). Jaipur, 2016: 1314 – 1318.

[13] Moody G B, Mark R G. The impact of the MIT-BIH arrhythmia database [J]. *IEEE Engineering in Medicine and Biology Magazine*, 2001, 20(3): 45 – 50.

[14] Blake C L, Merz C J. UCI repository of machine learning databases [EB/OL]. (2016 – 05 – 27). http://www.ics.uci.edu/~mlearn/MLRepository.html.

[15] Cover T M. Estimation by the nearest-neighbor rule [J]. *IEEE Transactions on Information Theory*, 1968, 14(1): 50 – 55.

[16] Dudani S A. The distance-weighted k-nearest-neighbor rule [J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976(4): 325 – 327.

[17] Zuo W, Zhang D, Wang K. On kernel difference-weighted k-nearest neighbor classification [J]. *Pattern Analysis and Applications*, 2008, 11(3/4): 247 – 257.

[18] Khare S, Bhandari A, Singh S, et al. ECG arrhythmia classification using spearman rank correlation and support vector machine [C]//Proceedings of the International Conference on Soft Computing for Problem Solving. Berlin: Springer, 2012: 591 – 598.

[19] Salzberg S L. On comparing classifiers: pitfalls to avoid and a recommended approach [J]. *Data Mining and Knowledge Discovery*, 1997, 1(3): 317 – 328.