计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220334
60(11): 2594−2610, 2023

基于异常特征模式的心电数据标签清洗方法

韩京宇　陈　伟　赵　静　郎　杭　毛　毅
（南京邮电大学计算机学院　南京　210023）
（江苏省大数据安全与智能处理重点实验室（南京邮电大学）　南京　210023）
（jyhan@njupt.edu.cn）

A Label Cleaning Method of ECG Data Based on Abnormality-Feature Patterns

Han Jingyu, Chen Wei, Zhao Jing, Lang Hang, and Mao Yi
（*School of Computer Science*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210023）
（*Jiangsu Key Laboratory of Big Data Security and Intelligent Processing* (*Nanjing University of Posts and Telecommunications*), *Nanjing* 210023）

Abstract　Automatic detection of electrocardiogram (ECG) abnormality is a typical multi-label classification problem, which heavily relies on sufficient samples with high-quality abnormality labels for model training. Unfortunately, we often face ECG datasets with partial and incorrect labels, so how to clean weakly-labelled datasets to obtain the clean datasets with all the correct abnormality labels is becoming a pressing concern. Under the assumption that we can have a small-sized example dataset with full and correct labels, we propose an abnormality-feature pattern (AFP) based method to automatically clean the weakly-labelled datasets, thus obtaining all the correct abnormality labels. The cleaning process proceeds with two stages, clustering-based rule construction and iteration-based label cleaning. During the first stage, we construct a set of label inclusion and exclusion rules and a set of binary discriminators by exploiting the different abnormality-feature patterns which are identified through Dirichlet process mixture model (DPMM) clustering. During the second stage, we first identify the relevant abnormalities according to the label inclusion and exclusion rules, and then refine the relevant abnormalities with iterations. AFP method takes advantage of the abnormality-feature patterns shared by the example dataset and weakly-labelled dataset, which is based on both the human intelligence and the correct label information from the weakly-labelled dataset. Further, the method stepwise removes the incorrect labels and fills in the missing ones with an iteration, thus ensuring a reliable cleaning process. The experiments on real and synthetic datasets prove the effectiveness of our method.

Key words　electrocardiogram (ECG); multi-label classification; abnormality labels; abnormality-feature pattern (AFP); binary discriminator; label cleaning

摘　要　心电图（electrocardiogram, ECG）异常的自动检测是一个典型的多标签分类问题，训练分类器需要大量有高质量标签的样本. 但心电数据集异常标签经常缺失或错误，如何清洗弱标签得到干净的心电数据集是一个亟待解决的问题. 在一个标签完整且准确的示例数据集辅助下，提出一种基于异常特征模式 (abnormality-feature pattern, AFP) 的方法对弱标签心电数据进行标签清洗，以获取所有正确的异常标签. 清洗分 2 个阶段，即基于聚类的规则构造和基于迭代的标签清洗. 在第 1 阶段，通过狄利克雷过程混合模型（Dirichlet process mixture model, DPMM）聚类，识别每个异常标签对应的不同特征模式，进而构建异常发现规则、排除规则和 1 组二分类器. 在第 2 阶段，根据发现和排除规则辨识初始相关标签集，然后根据

收稿日期：2022-04-25；**修回日期**：2022-12-09
基金项目：国家自然科学基金项目（62002174）
This work was supported by the National Natural Science Foundation of China (62002174).

二分类器迭代扩展相关标签并排除不相关标签. AFP 方法捕捉了示例数据集和弱标签数据集的共享特征模式，既应用了人的知识，又充分利用了正确标记的标签；同时，渐进地去除错误标签和填补缺失标签，保证了标签清洗的可靠性. 真实和模拟数据集上的实验证明了 AFP 方法的有效性.

关键词 心电图；多标签分类；异常标签；异常特征模式；二分类器；标签清洗

中图法分类号 TP391.5；TP181

根据世界卫生组织的报告，心血管疾病（cardiovascular diseases, CVDs）是人类健康的头号杀手[1]. 心电图（electrocardiogram, ECG）作为一种无创的心脏健康检测技术在临床上广泛应用，因而心电图异常的自动识别备受关注[2]. 由于每个样本通常会有多种心电异常，例如完全性左束支阻滞经常和前间壁心肌梗死共同出现，房性期前收缩经常和窦性心动过缓并发[3]，其自动检测在机器学习中是一个典型的多标签分类问题. 训练有效的分类器，通常需要大量具有完整且准确标签的样本，然而在实际中，人工标注心电异常不仅需要专业人员，而且费时费力.

众所周知，丰富且标记正确的样本对于训练有效的分类器至关重要，尤其是训练深度学习模型，样本的数量直接影响分类器的精度和泛化性. 现实中，经常有一些心电的弱标签数据集（weakly labelled dataset, *WD*）不能被有效利用[4-5]. 这些样本有异常标签，但标签不一定完整和正确，如何有效地去除错误标签、填补缺失标签，提供更丰富的训练数据集，意义重大[6]. 我们注意到，获取少量的有完整、正确标签的示例数据集（example dataset, *ED*）是完全可行的. 根据这个认识，弱标签心电数据集的清洗任务具体化为，给定一个 *WD* 和一个 *ED*，对 *WD* 中的异常标签进行清洗，获得干净数据集（clean dataset, *CD*），其每个样本有全部正确的异常标签.

目前关于弱标签心电图样本清洗的研究[7-10]可以分为 2 类：依赖于分类器的方法和独立于分类器的方法. 前者直接在弱标签数据集上训练一组分类器，并根据它们的判断来识别错误标记的样本[7,9-10]；后者旨在无需训练分类器的情况下，开发识别弱标签的专用算法[8]，本文提出的方法就属于后者. 另外，机器学习中利用弱标签数据和未标记数据训练通用分类器的方法也受到广泛关注. 前者根据每个样本的部分相关标签进行学习[11-15]，后者基于小部分正样本和大量未标记样本来训练分类器[16-17]. 但这些方法不能用于创建具有干净标签的数据，也不适用于心电图数据：首先，心电图异常标签多达几十个，而通常的方法只关注少量标签；其次，心电图数据的异常标签和特征间有复杂的相关性，即一个异常呈现不同的特征模式，而且类似的特征模式可能指示不同的异常[3,5].

本文提出一种基于异常特征模式（abnormality-feature pattern, AFP）清洗弱标签心电数据的方法（后文简称：AFP 方法），生成可重复使用的、具有完整且准确标签的干净数据集，为心电数据的有监督学习提供更丰富的训练样本. 具体地，基于心电数据的异常特征模式，在 *ED* 的支持下去除错误标签并填补缺失标签，AFP 包括 2 个阶段，即清洗规则构建和迭代清洗异常标签. 在第 1 阶段，提取由 *ED* 和 *WD* 共享的异常特征模式，识别 *ED* 和 *WD* 共享的异常标签，即锚标签；然后，提取异常发现规则、异常排除规则和 1 组二分类器. 在第 2 阶段，首先根据标签发现和排除规则识别初始相关异常；然后，根据二分类器迭代判断其他的弱标签是否属于对应样本. 迭代终止时，生成对应 *WD* 的 *CD*. 方法中 *ED* 的支持是不可或缺的，它不仅是发现共享异常特征模式的基础，也是挖掘清洗规则的源泉.

本文主要贡献在于提出了一种通用的心电数据标签清洗方法，具体包括 3 个方面：

1）提出利用异常特征模式识别以高置信度属于实例的锚标签. 既利用了人类知识，又提取了弱标签中的可靠信息，保证清洗方法的有效性和鲁棒性.

2）提出挖掘异常发现和排除规则的具体算法，这些规则是标签清洗的基础.

3）开发了一个迭代式的异常清洗框架，通过逐步缩小不确定区间来清洗异常标签，精准地去除错误标签和添加缺失标签，避免方法性能的波动.

1 相关工作

1.1 弱标签心电图数据学习

大多数心电图分类工作假设样本的标签完整且准确，但实际中很难满足. 在心电异常分类中，如何利用被错误标记的样本备受关注. 文献 [7] 中利用 5 种不同的分类器，支持向量机（support vector machine, SVM）、

K 近邻(K-nearest neighbor, KNN)、朴素贝叶斯(naive Bayesian, NB)、线性判别分析(linear discriminant analysis, LDA)和决策树(decision tree, DT)，将所有训练样本随机分成 10 份，1 份作为验证集，其余 9 份作为训练集，然后将训练集输入到这 5 种分类器，确定标签是否被错误标记. 文献 [8] 中自动删除具有潜在错误标签的训练样本，协助用户进行心电图病症标记. 所提出的方法基于遗传优化过程，其中每个染色体代表一个候选解决方案，用于确认无效的训练样本. 文献 [9] 中提出，用分类性能最好的前 k 个算法独立进行投票，如果 k 个算法对是否有某个标签持不同观点，则将该标签视为潜在错误标签.

另一项密切相关的工作是如何利用弱标签数据训练通用分类器[11-14,18-21]，该工作分成 2 类：

一类是直接对弱标签进行修正. 文献 [12] 中提出通过 10 折交叉验证来识别错误标记的数据，在第 i 轮中的第 i 组作为验证集，其余 9 组作为训练集. 在验证集上预测的每个标签，如果训练出来的分类器不一致，则被认为是不正确的标签. 文献 [13] 用一个矩阵建模图像和标签的相似性，通过矩阵补齐技术来补齐图像的标签，也是一种修补标签的方法. 文献 [18] 中提出了半监督弱标签(semi-supervised weak-label, SSWL)方法，解决基于部分标签甚至无标签数据进行的学习问题，它根据实例相似性和标签相似性来补充缺失标签.

另外一类是直接利用弱标签数据训练分类器. 文献 [20] 提出的随机梯度下降树(random gradient descent tree, RGD-tree)，在有错误标签的数据集上训练支持向量机，保证超平面的可分性. 文献 [21] 提出采用缩放铰链损失函数(rescaled hinge loss function)，提高支持向量机对噪声标签的鲁棒性.

本文不同于上述 2 类方法在于：1)标签被清洗后，数据集可以被复用于各种计算任务，不仅可以用于训练分类器，而且可以用于各种数据挖掘、数据分析任务，拓宽数据的可用性；2)弱标记数据进行学习通常只能在符合方法特点的数据集上进行有效学习，方法对数据集敏感，有一定的适用局限性.

1.2 PU 学习

另一个相关工作是正样本和未标记样本(positive and unlabeled, PU)学习[22]，它根据正样本和未标记样本来训练分类器，主要分为 2 步法(two-step methods)和有偏学习(biased learning).

2 步法的第 1 步识别出一些可靠的负样本；第 2 步将此样本与正样本结合用于训练分类器，对未识别的样本进行分类. 文献 [23] 基于正样本构建概率生成模型，把相对正例密度最低区域的样本认为是负样本，基于此构建分类模型. 文献 [24] 设计了最小平方支持向量机对未标记样本进行分类.

有偏学习训练分类器时将无标签样本当成负样本. 文献 [15] 提出采用多个合成器、过滤器和确认器标记无标签的样本. 文献 [16] 中将所有未标记的样本标记为负样本，并使用线性函数从噪声实例中学习，从而将问题转化为噪声学习问题. 文献 [17] 中引入了一种生成 PU 学习模型，在没有完全随机选择(selected completely at random, SCAR)假设的情况下，生成一组虚拟 PU 示例来训练分类器. 文献 [25] 将未标记的数据集视为负类，对负类标签进行建模，转化为使错误的负标签风险最小的问题.

1.3 多标签分类

因为分类器的输出空间大小与类标签数量成指数关系，所以多标签分类任务具有挑战性. 一般多标签分类可以通过 2 类方法来解决，即问题转换和算法适应[26-27]. 前者将多标签分类问题转化为其他成熟的学习场景，而后者采用流行的学习技术来处理多标签分类问题.

问题转换方法可以分为 3 类：二分类、标签排序和多类分类. 代表性的二分类有二元相关法[28]和分类器链法[29]，前者将多标签分类问题分解为 1 组独立的二分类问题，后者将多标签分类问题转化为二分类问题链，链中的后续二分类器建立在前面分类器的预测之上. 标签排序的代表是校准标签排名(calibrated label ranking, CLR)，它将多标签分类问题转化为标签排序问题，其中标签之间的排序通过成对比较来实现[30]. 诸如 Random K-Labelset[31] 之类的多类方法将多标签分类问题转换为多类分类问题的集合，其中每个组件分类器都针对标签的随机子集.

算法适应方法对已存算法进行改造实现多标签分类. 例如，文献 [32] 中的多标签 K 近邻(multi-label K-nearest neighbor, MLKNN)方法采用 K 近邻技术来处理多标签数据，使用最大后验(maximum a posteriori, MAP)规则进行预测. 多标签决策树(multi-label decision tree, ML-DT)采用决策树技术来处理多标签数据，基于多标签熵的信息增益标准递归地构建决策树[33]. 文献 [34] 提出的排序支持向量机(ranking support vector machine, Rank-SVM)采用最大边距策略进行多标签分类，优化了一组线性分类器以最小化经验排序损失. 文献 [35] 提出基于粒化特征加权的 K 近邻算法实现多标签学习.

1.4 噪声标签清洗

目前噪声标签清洗方法主要分成2类：一类对噪声鲁棒性进行建模，文献[36]提出对噪声的代理损失函数(surrogate loss function)和噪声率进行建模，文献[37]提出均匀标签噪声模型(uniform label noise model)，通过风险最小化，创建鲁棒性强的多标签分类模型. 另外一类基于模型过滤进行噪声标签清洗，如文献[38]提出基于数据分布过滤(data distribution filtering, DDF)的标签噪声过滤方法. 对于数据集中的每一个样本，根据其近邻内样本的分布，将其邻域样本形成的区域划分为高密度区域和低密度区域，然后针对不同的区域采用不同的噪声过滤规则进行过滤.

2 问题和方法概述

令$U=\{l^1,\cdots,l^k,\cdots,l^u\}$为所有异常标签，表1列出了一些常见的心电异常标签.$ED=\{ob_1,\cdots,ob_i,\cdots,ob_N\}$是具有正确标签的示例数据集，每个$ob_i$由特征$\boldsymbol{ft}(ob_i)$和相关异常标签集$rl(ob_i)\subseteq U$组成.$\boldsymbol{ft}(ob_i)$是一个$d$维向量$(f_1,\cdots,f_k,\cdots,f_d)$，每个$f_k$代表一个数值型特征，采用截断多元正态分布来描述该$d$维向量的分布. 本文中，对每个样本的心电数据经过波形去噪、波形(QRS波、P波、T波)识别、特征提取和归一化，在12个导联上提取横向间隔、纵向幅度、电轴倾斜和波形高度4类特征[39]，构成d维向量$(f_1,\cdots,f_k,\cdots,f_d)$. 给定一个待清洗的大型弱标签数据集$WD=\{ob_1,\cdots,ob_i,\cdots,ob_M\}$，每个$ob_i$带有弱标签集$cl(ob_i)$，$cl(ob_i)$中的一些标签属于相关标签集$rl(ob_i)$，而其余的则是错误标签，即不相关标签. 另外，$ob_i$的有些相关标签缺失. 清洗的目的是从$WD$生成一个$CD$. 下文除特殊说明，异常和标签指示同一概念.表2中列出了本文中使用的主要符号.

Table 1 Abnormality Labels in CHE and CHW Datasets

表1 CHE和CHW实验数据集中的异常标签

标签名	标签名
心房颤动	室性期前收缩
窦性心动过缓	交界性期前收缩
窦性心律不齐	左前分支阻滞
I度房室传导阻滞	左心室肥大
窦性心动过速	下壁心肌梗死
前间壁心肌梗死	完全性左束支阻滞
左心房肥大	不完全性右束支阻滞
完全性右束支阻滞	房性期前收缩

Table 2 Meanings of Key Notations in Our Paper

表2 本文中主要符号含义

符号	含义
$U=\{l^1,\cdots,l^k,\cdots,l^u\}$	所有异常(标签)
ob_j，f_k	实例，特征
$\boldsymbol{ft}(ob)$	实例ob的特征向量
ED，WD	示例数据集、弱标签数据集
CD	干净数据集
TD	ED和WD中锚标签样本形成的数据集
$cl(ob_i)$，$rl(ob_i)$	ob_i的弱标签集和相关标签集
$al(ob)$	实例ob的锚标签集
$\overline{ED}(l)$，$\underline{ED}(l)$	ED中含和不含标签l的样本
$\overline{WD}(l)$，$\underline{WD}(l)$	WD中含和不含标签l的样本
$\overline{FC}(l)$，$\underline{FC}(l)$	标签l的正样本和负样本上的所有类簇
$\overline{C}_i(l)$，$\underline{C}_i(l)$	标签l的正样本和负样本上的第i个类簇
$fp_j(l)$	l对应的第j个异常特征模式
$\overline{FP}(l)$，$FP(l)$	l正样本和负样本上异常特征模式集
$fq(l)$	标签l在数据集上的出现次数
$AWD(FP^1,FP^2)$	异常特征模式集合FP^1，FP^2的平均Wasserstein距离
$supp$，$conf$，$cort$	支持度、置信度和正相关度
st，ct，rt	支持度、置信度和正相关度的阈值
$dr(ob,l)$	标签l属于实例ob的判别比
θ^l	标签l属于实例的分割阈值
ρ^l	标签l属于实例的模糊间隔长度
$lf(ob)$	实例ob的生存指数

给定一个标签$l\in U$，它在ED(或WD)上的正样本集，用$\overline{ED}(l)$(或$\overline{WD}(l)$)表示，是ED(或WD)上有l的样本集；它在ED(或WD)上的负样本集，用$\underline{ED}(l)$(或$\underline{WD}(l)$)表示，是ED(或WD)上没有l的样本集. 提出的异常特征模式方法，利用异常对应的特征模式以及异常间的关系来修复标签. 具体来说，每个异常标签l在ED的正样本上对应一组类簇$\overline{FC}^{ED}(l)=\{\overline{C}_1(l),\overline{C}_2(l)\cdots,\overline{C}_i(l),\cdots,\overline{C}_n(l)\}$，在$ED$的负样本上对应一组类簇$\underline{FC}^{ED}(l)=\{\underline{C}_1(l),\underline{C}_2(l),\cdots,\underline{C}_i(l),\cdots,\underline{C}_m(l)\}$. 类似地，生成$\overline{FC}^{WD}(l)$和$\underline{FC}^{WD}(l)$. 相应地，每个异常$l$在正样本和负样本上各对应一组异常特征模式集$FP(l)$，定义如下.

定义1. 异常特征模式. 给定l的一个类簇$C_i(l)$，它的异常特征模式$fp_i(l)$对应一个截断多元正态分布$NM(\boldsymbol{\mu}_i(l),\boldsymbol{\Sigma}_i(l))$，其中$\boldsymbol{\mu}_i(l)$是特征均值，$\boldsymbol{\Sigma}_i(l)$是特征协方差.

给定2个特征模式$fp_1=NM(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$和$fp_2=NM(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$，衡量$fp_1$和$fp_2$的相异性Wasserstein距离为：

$$wdt(fp_1,fp_2)=\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2^2+\mathrm{tr}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2(\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{\frac{1}{2}})^{\frac{1}{2}}),\tag{1}$$

其中$\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2$是 L2 范数距离.

AFP 方法分成 2 个阶段:基于聚类的清洗规则构造和基于迭代的标签清洗,如图 1 所示. 在基于聚类的清洗器构造时,首先在 *ED* 和 *WD* 上进行聚类寻找锚模式.

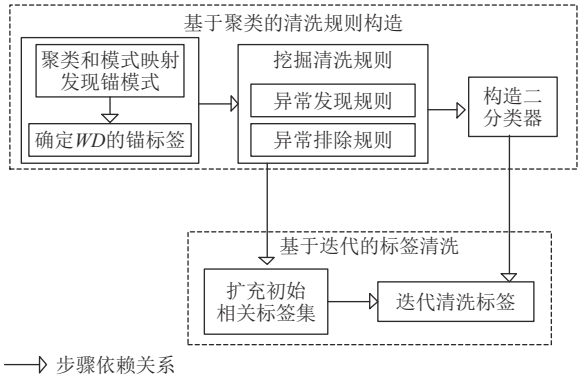

Fig. 1 The steps of AFP

图 1 AFP 方法的步骤

定义 2. 锚模式. 给定一个异常特征模式 $fp_i(l)$,如果它被 *ED* 和 *WD* 共享,它就是一个锚模式.

定义 3. 锚异常集. 给定一个实例 $ob\in WD$,其锚异常集 $al(ob)\subseteq rl(ob)$是根据锚模式识别的相关异常集.

锚异常集是根据共享的锚模式识别出的 *WD* 上的高置信度标签,它既是 *WD* 相关标签的一部分,又用来扩充规则挖掘依赖的样本.

挖掘标签发现规则和标签排除规则,分别用来表征 2 个异常特征模式的正相关性和负相关性,在后续的标签清洗中分别用于填补缺失标签和去除错误标签. 最后,为每个异常构造二分类器,以支持后续的标签迭代清洗.

标签迭代清洗前,在 *ED* 和 *WD* 组成的 *TD* 上构建隔离森林 *iForest*(isolation forest)[40],根据样本在隔离森林中的路径长度决定参与迭代清洗的次数. 清洗时,首先根据标签发现和排除规则,包含或排除弱标签,包含的标签确定为相关标签,排除的标签视为不相关标签,从而扩充初始相关标签集,缩小了弱标签集的大小;然后根据二分器,迭代清洗弱标签集,逐步缩小不确定的标签集合. 迭代清洗时,通过不断地逼近标签和类簇特征间的关联,识别出其他相关标签.

后文除特别说明,使用 Jensen-Shannon 距离来衡量 2 个分布的差异,记为 *JSD*.

定义 4. *JSD*. 给定 2 个分布 $\boldsymbol{P}(X)$和 $\boldsymbol{Q}(X)$,其中 X 表示域值,其 *JSD* 定义为

$$JSD(\boldsymbol{P}\|\boldsymbol{Q})=\frac{1}{2}(D(\boldsymbol{P}\|\mathbf{M})+D(\boldsymbol{Q}\|\mathbf{M})), \tag{2}$$

其中 $\boldsymbol{M}=\frac{1}{2}(\boldsymbol{P}+\boldsymbol{Q})$, $D(\boldsymbol{P}\|\mathbf{M})$是 $\boldsymbol{P}$ 和 $\boldsymbol{M}$ 之间的相对熵, $D(\boldsymbol{Q}\|\mathbf{M})$是 $\boldsymbol{Q}$ 和 $\boldsymbol{M}$ 之间的相对熵.

3 标签清洗规则的构造

对于每个异常 l,在其正样本和负样本上分别识别一组类簇,进而构建 l 对应的 1 组特征模式. 虽然每个样本表征为高维数据,但本文没有对数据进行降维处理,因为有些心电病症的区别主要集中在若干特征上[3],如果进行降维,会剔除或淹没这些关键信息,降低异常识别精度. 对样本进行聚类时,没有采用常见的方法如 *k*-均值(*k*-Means)进行聚类,避免根据经验指定类簇数量,而是采用狄利克雷过程混合模型(Dirichlet process mixture model, DPMM)进行聚类,它能够自适应地根据数据分布特点发现最合适的类簇[41].DPMM 中每个实例 ob_i 产生于中国餐馆过程 *CRP*(Chinese restaurant process)[42] 表达的狄利克雷过程:

$$(\boldsymbol{\mu},\boldsymbol{\Sigma})_{1,2,\cdots,\infty}\sim NIW(\boldsymbol{\mu}_0,k_0,\boldsymbol{Y}_0,v_0), \tag{3}$$

$$Z_{1,2,\cdots,i,\cdots,N}\sim CRP(\gamma), \tag{4}$$

$$ob_{1,2,\cdots,i,\cdots,N}\sim MN((\boldsymbol{\mu},\boldsymbol{\Sigma})_{Z_i}). \tag{5}$$

该生成模型中,实例由多元正态分布 *MN* 产生,类簇分配由中国餐馆过程 $CRP(\gamma)$ 决定,其中 γ 是聚焦参数, Z_i 是实例 ob_i 对应的类簇;作为狄利克雷过程的基分布,逆威沙特分布 *NIW*(normal-inverse-Wishart)是多元正态分布 *MN* 的共轭先验分布: $\boldsymbol{\mu}_0$是 N 维向量,代表最初平均值; k_0 用作平滑因子,控制 $\boldsymbol{Y}_0$中各个元素的放缩比例; v_0 是自由度,初始化为原始特征数目; $\boldsymbol{Y}_0$是成对偏差积,初始化为 $N\times N$ 的常数矩阵.

为了找到实例 ob_i 所属合适类簇,算法 1 用吉布斯采样获得类簇分配.

算法 1. clusterAssignment.

输入:训练集 $D=\{ob_1,ob_2,\cdots,ob_N\}$, DPMM 参数;

输出:每个实例 $ob_i(1\leqslant i\leqslant N)$的类簇分配.

① 初始化参数 $Z_i(1\leqslant i\leqslant N)$和循环变量 t;
② while t do
③ for $i\leftarrow 1$ to N do
④ 根据式(6)计算 ob_i 第 t 轮的类簇分配概率;
⑤ end for
⑥ if $(Z_1^{(t)},Z_2^{(t)},\cdots,Z_N^{(t)})$和$(Z_1^{(t-1)},Z_2^{(t-1)},\cdots,Z_N^{(t-1)})$相同 then
⑦ $t\leftarrow 0$;
⑧ end if
⑨ t++;
⑩ end while

⑪ return $\{Z_1,Z_2,\cdots,Z_N\}$.

类簇分配不停迭代，直到不再改变. 迭代时，每个实例的类簇分配概率根据式(6)更新：

$$\begin{aligned}&P(Z_i=m|Z_{-i},ob_{1:i})\propto\\&P(Z_i=m|Z_{-i})P(ob_i|Z_i=m,Z_{-i}),\end{aligned}\tag{6}$$

其中 Z_{-i} 是除 ob_i 之外的所有实例的类簇分配.

证明.

$$P(Z_i=m|Z_{-i},ob_{1:i})=\frac{P(Z_i=m,Z_{-i},ob_{1:i})}{P(Z_{-i},ob_{1:i})}=$$

$$\frac{P(Z_{-i})P(Z_i=m|Z_{-i})P(ob_{-i}|Z_i=m,Z_{-i})P(ob_i|ob_{-i},Z_i=m,Z_{-i})}{P(Z_{-i})P(ob_{-i}|Z_{-i})P(ob_i|ob_{-i},Z_{-i})}$$

$$=\frac{P(Z_i=m|Z_{-i})P(ob_{-i}|Z_i=m,Z_{-i})P(ob_i|ob_{-i},Z_i=m,Z_{-i})}{P(ob_{-i}|Z_{-i})P(ob_i|ob_{-i},Z_{-i})}.$$

由于$P(ob_{-i}|Z_i=m,Z_{-i})=P(ob_{-i}|Z_{-i})$和

$$P(ob_i|ob_{-i},Z_i=m,Z_{-i})=P(ob_i|Z_i=m,Z_{-i}),$$

可得

$$\begin{aligned}&P(Z_i=m|Z_{-i},ob_{1:i})\propto\\&P(Z_i=m|Z_{-i})P(ob_i|Z_i=m,Z_{-i}).\end{aligned}$$

证毕.

式(6)第 2 行的第 1 项是给定 ob_i 之外的所有实例的类簇分配条件下 ob_i 的类簇分配，根据式(7)的中国餐馆过程来确定：

$$P(Z_i=m|Z_{-i})=\begin{cases}\dfrac{n_{m,-i}}{i+\gamma-1},m\text{是现存簇},\\[2ex]\dfrac{\gamma}{i+\gamma-1},m\text{是新簇},\end{cases}\tag{7}$$

其中 $n_{m,-i}$ 是簇 m 中除 ob_i 外的实例数.

式(6)第 2 行的第 2 项是给定当前所有类簇分配条件下 ob_i 的概率，根据多元正态分布确定：

$$P(ob_i|Z_i=m,Z_{-i})\propto MN(\boldsymbol{\mu}_{m,-i},\boldsymbol{\Sigma}_{m,-i}),\tag{8}$$

其中 $\boldsymbol{\mu}_{m,-i}$ 和 $\boldsymbol{\Sigma}_{m,-i}$ 是类簇 m 不包括 ob_i 时的均值和协方差.

3.1 基于异常特征模式识别锚异常

为了识别 *WD* 中实例的锚异常，首先识别 *ED* 和 *WD* 共享的异常特征模式. 给定一个标签 l, *ED* 和 *WD* 的正例上对应的异常特征模式集记为 $\overline{FP}^{ED}(l)=\left\{\overline{fp}_1^{ED}(l),\overline{fp}_2^{ED}(l),\cdots,\overline{fp}_i^{ED}(l),\cdots,\overline{fp}_n^{ED}(l)\right\}$和 $\overline{FP}^{WD}(l)=\left\{\overline{fp}_1^{WD}(l),\overline{fp}_2^{WD}(l),\cdots,\overline{fp}_i^{WD}(l),\cdots,\overline{fp}_m^{WD}(l)\right\}$，如果$\left|\overline{FP}^{ED}(l)\right|$与$\left|\overline{FP}^{WD}(l)\right|$不同，在较大的集合中删除那些与较小集合中各个模式差异最大的那些模式，使 2 个特征模式集合大小一样.

如果 *ED* 和 *WD* 共享某个锚模式，则在 2 个数据集上对应的模式表达不仅应该相似，而且对应的 2 个类簇上应有尽可能多的具有相同标签集 $ls\in 2^U$ 的样本. 因此，2 个模式集合的最优一对一映射

$$f:\overline{FP}^{ED}(l)\leftrightarrow\overline{FP}^{WD}(l)\tag{9}$$

要满足 2 个条件：

1)配对的异常特征模式的平均 Wasserstein 距离

$$AWD\left(\overline{FP}^{ED}(l),\overline{FP}^{WD}(l)\right)=\frac{\sum_{i=1}^{k}wdt\left(\overline{fp}_i^{ED}(l),\overline{fp}_i^{WD}(l)\right)}{k\cdot z}\tag{10}$$

应最小化，其中 k 是 $\overline{FP}^{ED}(l)$的特征数量，z 是确保 $AWD\left(\overline{FP}^{ED}(l),\overline{FP}^{WD}(l)\right)$介于 0~1 之间的规范化因子，$wdt\left(\overline{fp}_i^{ED}(l),\overline{fp}_i^{WD}(l)\right)$是 $\overline{fp}_i^{ED}(l)$和 $\overline{fp}_i^{WD}(l)$的 Wasserstein 距离.

2)配对的异常特征模式共享尽可能多地具有相同标签集的样本. 给定 2 个类簇$C_i^{ED}(l)$和$C_i^{WD}(l)$, 用$fq(ls)$表示标签集 ls 在类簇中的出现频率. 设$C_i^{ED}(l)$和$C_i^{WD}(l)$的标签集分布分别是$md\left(C_i^{ED}(l)\right)=(fq(ls_1),fq(ls_2),\cdots,fq(ls_{|2^U|}))$和 $md(C_i^{WD}(l))=(fq(ls_1),\ fq(ls_2),\cdots,\ fq(ls_{|2^U|}))$. 因此，要最小化式(11)：

$$AJSD\left(\overline{FP}^{ED}(l),\overline{FP}^{WD}(l)\right)=\frac{1}{k}\sum_{i=1}^{k}JSD(md(C_i^{ED}(l)),md(C_i^{WD}(l))).\tag{11}$$

给定$\overline{FP}^{ED}(l)$中各个模式的排序，需要找到满足上述 2 个条件的$\overline{FP}^{WD}(l)$的对应元素排列. 这是一个多目标优化问题，采用模拟退火[43]寻找非劣解，算法流程如图 2 所示. 模拟退火由 3 个参数控制：初始温度 tl、降温速率 $cr(0<cr<1)$和候选解个数 ss.2 维矩阵 $\boldsymbol{F}[ss,k]$ 的每一行代表 l 在 *WD* 上各个异常特征模式的一个排列. 随着 tl 下降，在每个温度，算法为每个候选序列生成一个新排列，并将其与现存排列进行比较. 如果新排列的收益大于现存排列，则现存排列被取代；否则，根据概率替换. 收益定义为式(12)：

$$ben\left(\overline{FP}_{\text{new}}^{WD}(l),\overline{FP}_{\text{old}}^{WD}(l)\right)=\begin{cases}1&,dt^{AWD}\leqslant 0\text{且}dt^{AJSD}\leqslant 0,\\-|dt^{AWD}+dt^{AJSD}|&,\text{其他}.\end{cases}\tag{12}$$

其中

$$\begin{aligned}dt^{AWD}&=AWD\left(\overline{FP}^{ED}(l),\overline{FP}_{\text{new}}^{WD}(l)\right)-AWD\left(\overline{FP}^{ED}(l),\overline{FP}_{\text{old}}^{WD}(l)\right),\\dt^{AJSD}&=AJSD\left(\overline{FP}^{ED}(l),\overline{FP}_{\text{new}}^{WD}(l)\right)-AJSD\left(\overline{FP}^{ED}(l),\overline{FP}_{\text{old}}^{WD}(l)\right).\end{aligned}$$

式(12)的直观含义:如果新排列优于原排列，则返回 1，否则返回$-|dt^{AWD}+dt^{AJSD}|$. 当迭代结束时，有 ss 个候选解，从候选解中选择一个最优解或非劣解. 模式排列算法的运行时间主要由嵌套循环决定，其时间复杂

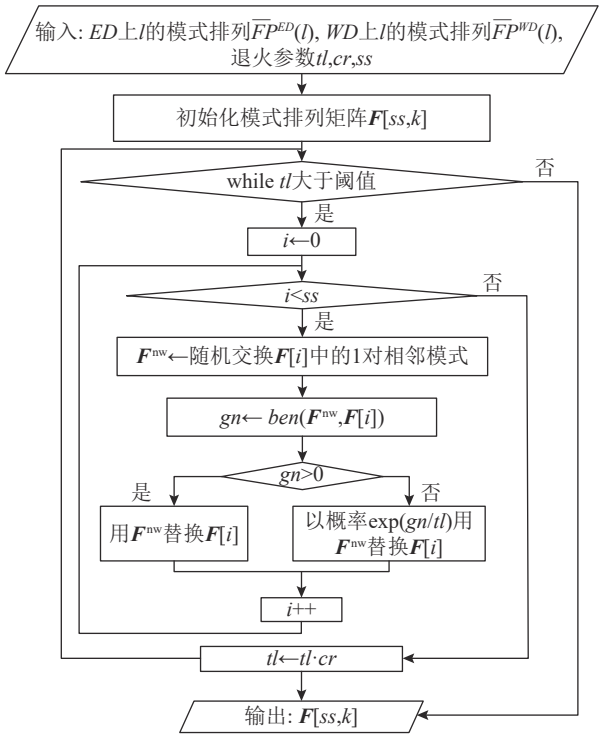

Fig. 2 Flow chart of algorithm for pattern ordering on *WD*

图 2 *WD* 上模式排序算法流程图

度为$O(ss\cdot\log_{cr}tl)$.

最后, 模式排列算法返回的每个候选解, 对应 1 组 *AWD* 和 *AJSD*, 计算这 *ss* 个候选解的平均值作为阈值. 然后, 将平均值低于该阈值的候选解作为锚模式. 一旦确定了标签的锚模式, 给定一个实例 $ob\in WD$, 锚模式对应的异常标签称为 *ob* 的锚标签, 同时是该实例的相关标签. 然后, 将 *ED* 和 *WD* 的锚标签样本结合, 形成一个训练数据集 *TD*, 在 *TD* 上挖掘标签发现、排除规则并构建二分类器.

3.2 挖掘标签发现和排除规则

3.2.1 在 *TD* 上挖掘标签发现规则

在心电数据中, 一个异常经常表现出若干特征模式. 标签发现规则用来指示频繁共同出现的异常特征模式. 给定 2 个异常特征模式 $fp_i(l^s)$和 $fp_j(l^t)$, 标签发现规则 $fp_i(l^s)$ => $fp_j(l^t)$表明: 若某个实例同时落入 fp_i 和 fp_j 的特征模式, 并且该实例有异常标签 l^s, 则该实例有异常标签 l^t.

例 1. 设有 2 个异常标签 *A* 和 *B*, *A* 是前壁心肌梗死, *B* 是左后分支传导阻滞. 假设某个标签发现规则是 $fp_i(A)$=>$fp_j(B)$, 其中

$$fp_i(A)=NM\left((0.75,0.83),\begin{pmatrix}0.3&0.15\\0.15&0.2\end{pmatrix}\right),$$

$$fp_j(B)=NM\left((0.76,0.81),\begin{pmatrix}0.3&0.6\\0.6&0.4\end{pmatrix}\right).$$

这表明特征模式为 $NM\left((0.75,0.83),\begin{pmatrix}0.3&0.15\\0.15&0.2\end{pmatrix}\right)$ 的前壁心肌梗死频繁与特征模式是 $NM\left((0.76,0.81),\begin{pmatrix}0.3&0.6\\0.6&0.4\end{pmatrix}\right)$的左后分支传导阻滞共同呈现. 假设某个实例 *ob* 具有特征 $\boldsymbol{ft}(ob)=(0.755,0.82)$和标签 *A*, 可以推断 *ob* 也有标签 *B*, 因为 $\boldsymbol{ft}(ob)$同时呈现 $fp_i(A)$和 $fp_j(B)$这 2 个异常特征模式.

令 $fq(fp_i(l^s))$和 $fq(fp_j(l^t))$分别代表 $C_i(l^s)$和 $C_j(l^t)$中呈现 $fp_i(l^s)$和 $fp_j(l^t)$模式的样本个数, 则标签发现规则的支持度和置信度定义为式(13)(14):

$$supp(fp_i(l^s)\Rightarrow fp_j(l^t))=fq(fp_i(l^s)\cup fp_j(l^t)),\tag{13}$$

其中 $fq(fp_i(l^s)\cup fp_j(l^t))$是同时呈现 $fp_i(l^s)$和 $fp_j(l^t)$模式的样本个数.

$$conf(fp_i(l^s)\Rightarrow fp_j(l^t))=\frac{fq(fp_i(l^s)\cup fp_j(l^t))}{fq((fp_i(l^s))}.\tag{14}$$

模式的正相关性根据 Kulczynski(记为 *Kulc*)度量:

$$cort(fp_i(l^s)\Rightarrow fp_j(l^t))=Kulc(fp_i(l^s),fp_j(l^t))=\frac{1}{2}\left(\frac{fq(fp_i(l^s)\cup fp_j(l^t))}{fq(fp_i(l^s))}+\frac{fq(fp_i(l^s)\cup fp_j(l^t))}{fq(fp_j(l^t))}\right).\tag{15}$$

直观地, 如果 *cort*=0.5, 则 $fp_i(l^s)$和 $fp_j(l^t)$相互独立; 如果 *cort* 接近 1, 则 $fp_i(l^s)$和 $fp_j(l^t)$正相关; 如果 *cort* 接近 0, 则 $fp_i(l^s)$和 $fp_j(l^t)$呈负相关.

综上所述, 给定支持度阈值 *st*、置信度阈值 *ct* 和正相关阈值 *rt*, 一个标签发现规则 $fp_i(l^s)$ => $fp_j(l^t)$必须满足 3 个条件: 1)$supp(fp_i(l^s)$ => $fp_j(l^t))\geqslant st$; 2)$conf(fp_i(l^s)$ => $fp_j(l^t))\geqslant ct$; 3)$cort(fp_i(l^s)$ => $fp_j(l^t))\geqslant rt$.

本文通过 2 个步骤挖掘标签发现规则. 首先, 根据支持度阈值 *st* 和置信度阈值 *ct*, 挖掘两两标签间的关联规则[44]. 每个关联规则 $l^s\rightarrow l^t$表明, 如果 l^s出现, 则 l^t就会出现. 进一步, 根据算法 2 将关联规则提炼为标签发现规则.

算法 2. generateInclusionRule.

输入: 标签间的关联规则 *LR*, 异常特征模式 *AFP*, 支持度阈值 *st*, 置信度阈值 *ct*, 正相关阈值 *rt*;

输出: 标签发现规则.

① 初始化;

② foreach $(l^s\rightarrow l^t)\in LR$ do

③ L^{FP}←在 *AFP* 发现 l^s的所有异常特征模式;

④ R^{FP}←在 *AFP* 发现 l^t的所有异常特征模式;

/*下面根据异常模式对应的标签频率挖掘规则*/

⑤ foreach $(fp_i(l^s),fp_j(l^t))\in(L^{FP}\times R^{FP})$ do

⑥ if $(fp_i(l^s),fp_j(l^t))$满足式(13)~(15) then

⑦ put $fp_i(l^s)$=>$fp_j(l^t)$ to *ret*;

⑧ end if

⑨ end for

⑩ end for

⑪ return *ret*.

算法 2 的时间复杂度为 $O(|LR|\cdot q^2)$，其中$|LR|$是标签间的关联规则数量，q 是一个异常对应的特征模式数量的上界.

3.2.2 在 *TD* 上挖掘标签排除规则

定义 5. 标签排除规则. 给定 2 个异常标签l^s和l^t，如果

$$fq(l^s\cup l^t)\ll fq(l^s)\cdot fq(l^t), \tag{16}$$

则认为l^s和l^t是强负相关的，记$l^s \Leftrightarrow l^t$，其中 $fq(l^s\cup l^t)$是同时有标签l^s和l^t的样本个数.

为了度量强负相关性，引入阈值 $\varepsilon(0<\varepsilon\ll 1)$，如果

$$\frac{fq(l^s\cup l^t)}{fq(l^s)\cdot fq(l^t)}<\varepsilon, \tag{17}$$

则认为 $fp(l^s)$和 $fp(l^t)$是强负相关的. 直观含义是，如果l^s在某个实例上呈现，则l^t不会在该实例呈现，反之亦然. 采用算法 3 实现标签排除规则的挖掘.

算法 3. generateExclusionRule.

输入：频繁标签对集合 *FS*，负相关阈值 ε;

输出：标签排除规则.

① *ret*←Ø;

② foreach$\langle l^s, l^t\rangle \in FS$ do

③ if $\frac{fq(l^s\cup l^t)}{fq(l^s)\cdot fq(l^t)}<\varepsilon$ then

④ put $l^s \Leftrightarrow l^t$ to *ret*;

⑤ end if

⑥ end for

⑦ return *ret*.

算法 3 的时间复杂度为 $O(|FS|)$，其中$|FS|$是频繁标签对的数量.

3.3 构造二分类器

对于每个异常 *l*，通过在 *TD* 上聚类，分别得到其正例类簇 $\overline{FC}^{TD}(l)=\left\{\overline{C}_1^{TD}(l),\overline{C}_2^{TD}(l),\cdots,\overline{C}_i^{TD}(l),\cdots,\overline{C}_n^{TD}(l)\right\}$和负例类簇 $\underline{FC}^{TD}(l)=\left\{\underline{C}_1^{TD}(l),\underline{C}_2^{TD}(l),\cdots,\underline{C}_j^{TD}(l),\cdots,\underline{C}_m^{TD}(l)\right\}$，相应的异常特征模式分别是 $\overline{FP}^{TD}(l)=\left\{\overline{fp}_1^{TD}(l),\overline{fp}_2^{TD}(l),\cdots,\overline{fp}_i^{TD}(l),\cdots,\overline{fp}_n^{TD}(l)\right\}$和 $\underline{FP}^{TD}(l)=\left\{\underline{fp}_1^{TD}(l),\underline{fp}_2^{TD}(l),\cdots,\underline{fp}_j^{TD}(l),\cdots,\underline{fp}_m^{TD}(l)\right\}$.

给定实例 *ob* 和异常 *l*，计算 *ob* 和 *l* 的所有类簇中心的最小 Jensen-Shannon 距离. 假设 $\overline{FC}^{TD}(l)$各类簇中心是$\left(\overline{clc}_1^{TD}(l),\overline{clc}_2^{TD}(l),\cdots,\overline{clc}_i^{TD}(l),\cdots,\overline{clc}_n^{TD}(l)\right)$，$\underline{FC}^{TD}(l)$的各类簇中心是$\left(\underline{clc}_1^{TD}(l),\underline{clc}_2^{TD}(l),\cdots,\underline{clc}_j^{TD}(l),\cdots,\underline{clc}_m^{TD}(l)\right)$，所以 *ob* 和$\overline{FC}^{TD}(l)$的最小距离为

$$mind\left(ob,\overline{FC}^{TD}(l)\right)=\min\left\{JSD(\boldsymbol{ft}(ob),\overline{clc}_i^{TD}(l))|1\leqslant i\leqslant n\right\}, \tag{18}$$

ob 和 $\underline{FC}^{TD}(l)$的最小距离为

$$mind\left(ob,\underline{FC}^{TD}(l)\right)=\min\left\{JSD\left(\boldsymbol{ft}(ob),\underline{clc}_j^{TD}(l)\right)|1\leqslant j\leqslant m\right\}. \tag{19}$$

那么，*l* 关于 *ob* 的判别比为

$$dr(ob,l)=\frac{mind\left(ob,\overline{FC}^{TD}(l)\right)}{mind\left(ob,\overline{FC}^{TD}(l)\right)+mind\left(ob,\underline{FC}^{TD}(l)\right)}. \tag{20}$$

标签清洗中，根据式(21)判断 *l* 是否属于 *ob*：

$$hasLabel(dr,\theta^l,\rho^l)=\begin{cases}1, & dr\geqslant\theta^l+\rho^l,\\ -1, & dr\leqslant\theta^l-\rho^l,\\ 0, & \text{其他}.\end{cases} \tag{21}$$

这里 θ^l 是分割阈值，介于 0~1 之间，ρ^l 是模糊间隔长度. 如果 *hasLabel* 返回 1，*l* 是 *ob* 的相关标签；如果返回−1，*l* 是 *ob* 的无关标签；否则，无法确定 *l* 是否属于 *ob*，需要在下一轮迭代判断. 因为 *dr* 的值介于 0~1 之间，所以符合 Beta 分布：

$$f(dr:\alpha,\beta)=\frac{1}{B(\alpha,\beta)}dr^{\alpha-1}(1-dr)^{\beta-1}, \tag{22}$$

其中 α,β 是确定密度函数形状的参数. 则平均值μ^*和标准差δ分别是

$$\mu^*=\frac{\alpha}{\alpha+\beta}, \tag{23}$$

$$\delta=\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}. \tag{24}$$

因此，设置$\theta^l=\mu^*$，$\rho^l=\delta$.

最后，为所有标签构造二分类器，表示为 $BD=\left\{(\overline{FC}(l),\underline{FC}(l),\overline{FP}(l),\underline{FP}(l),\alpha_l,\beta_l)|1\leqslant l\leqslant u^*\right\}$.

4 迭代清洗 *WD* 中的弱标签

WD 中的异常标签通过 2 个步骤进行清洗，即弱标签预处理和迭代清洗.

4.1 弱标签预处理

给定一个实例 $ob\in WD$，其锚标签集合 $al(ob)$属于相关标签集合 $rl(ob)$.$cl(ob)$代表弱标签集合，其中的标签可能属于 $rl(ob)$，也可能不属于 $rl(ob)$. 对弱标签预处理时，确认弱标签是相关或不相关标签，从而缩小弱标签集合. 具体过程如算法 4 所示，给定一个实例 *ob*，如果它落入一个标签发现规则两侧的异常特征模式，并具有该规则的左侧标签，则右侧标签属于其相关异常 $rl(ob)$. 具体地，给定实例 *ob* 和异常特征模式 $NM(\boldsymbol{\mu},\boldsymbol{\Sigma})$，如果 *ob* 落入$(\boldsymbol{\mu}-3\cdot\boldsymbol{\Sigma},\boldsymbol{\mu}+3\cdot\boldsymbol{\Sigma})$区

间, 则 $ob \in NM$. 对不相关标签排除时, 如果标签排除规则一侧的标签属于给定实例, 则丢弃另一侧的标签, 将该标签从弱标签集合 $cl(ob)$ 中删除.

算法 4. reduceWeakLabelSet.

输入: 样本 ob, 标签发现规则 IR, 标签排除规则 ER;

输出: ob 的相关标签和缩减后的弱标签集合.

① $rl(ob) \leftarrow al(ob)$;

② $cl(ob) \leftarrow cl(ob) \backslash al(ob)$;

/*下面识别相关标签*/

③ foreach $l \in cl(ob)$ do

④ foreach $ir \in IR$ do

⑤ if ($ir.left.label \in rl(ob)$) $\wedge$ ($ir.right.label=l$) $\wedge$ ($ob \in ir.left.NM$) $\wedge$ ($ob \in ir.right.NM$) then

/*下面分别加入和排除标签*/

⑥ $rl(ob) \leftarrow rl(ob) \cup \{l\}, cl(ob) \leftarrow cl(ob) \backslash \{l\}$;

⑦ end if

⑧ end for

⑨ end for

/*下面排除不相关标签*/

⑩ foreach $l \in cl(ob)$ do

⑪ foreach $er \in ER$ do

⑫ if ($er.left.label \in rl(ob) \wedge er.right.label=l$) or ($er.right.label \in rl(ob) \wedge er.left.label=l$) then

⑬ $cl(ob) \leftarrow cl(ob) \backslash \{l\}$; /*排除标签*/

⑭ end if

⑮ end for

⑯ end for

⑰ return $rl(ob), cl(ob)$.

算法 4 的运行时间取决于一个实例的标签数和针对一个标签的规则数, 所以它的时间复杂度是 $O(u \cdot M^r)$, 其中 u 是 U 的大小, M^r 是单个异常的标签发现或排除规则的最大数目. 实际中, 算法 4 运行时间远小于 $O(u \cdot M^r)$, 因为一个实例的标签数目通常远小于 u.

4.2 迭代清洗弱标签

标签清洗时, 二分类器 BD 迭代地对剩余的弱标签进行区分, 扩展相关标签集合或从 $cl(ob)$ 中清除不相关标签, 同时更新二分类器 BD. 为避免 ob 无休止地参与迭代, 须设定其生存指数 $lf(ob)$. 为此, 在 $ED \cup WD$ 上构建隔离森林 $iForest$[40]. 实例在隔离森林中的平均路径长度 $apl(ob)$ 作为 ob 的生存指数分量. 每轮迭代中, ob 的生存指数 $lf(ob)$ 修改为:

$$lf(ob)=\begin{cases}\dfrac{x\cdot(|cl(ob)|+1)}{apl(ob)}, & |cl(ob)|\text{改变},\\ lf(ob)-1, & \text{否则},\end{cases} \tag{25}$$

其中 x 是控制变化率的因子. 式(25)的合理性在于, $apl(ob)$ 越大, ob 越可能被经常出现的特征模式覆盖, 因此需要的迭代次数越少; $|cl(ob)|$ 越大, 需要越多的迭代来区分其中的相关标签和非相关标签.

迭代清洗的算法流程如图 3 所示, 迭代直到所有弱标签被分类为相关标签或不相关标签, 或生存指数小于等于 0. 在 ob 到期后, 如果仍无法确定标签 l 是否属于 ob, 将这项任务留给人工识别. 每轮循环时, 一方面确定相关和不相关标签, 另一方面调用 *updateDiscriminator* 更新异常特征模式参数和所有标签的二分类器. 迭代清洗算法的时间复杂度是 $O(N \cdot u^l \cdot lf^m)$, 其中 N 是 WD 的大小, u^l 是一个实例的弱标签数目的上界, lf^m 是实例的生命周期的上界.

图 3 中的标签迭代清洗调用 *updateDiscriminator* 实现二分类器更新, 二分类器更新的算法流程如图 4 所示. 首先, 将新识别的实例和标签分配给相应的正、负类簇, 并调整 Beta 分布, 进而根据类簇样本调整异常特征模式参数. 这是为每个异常标签 l 调整分割阈值 θ^l 和模糊区间 ρ^l 的基础.

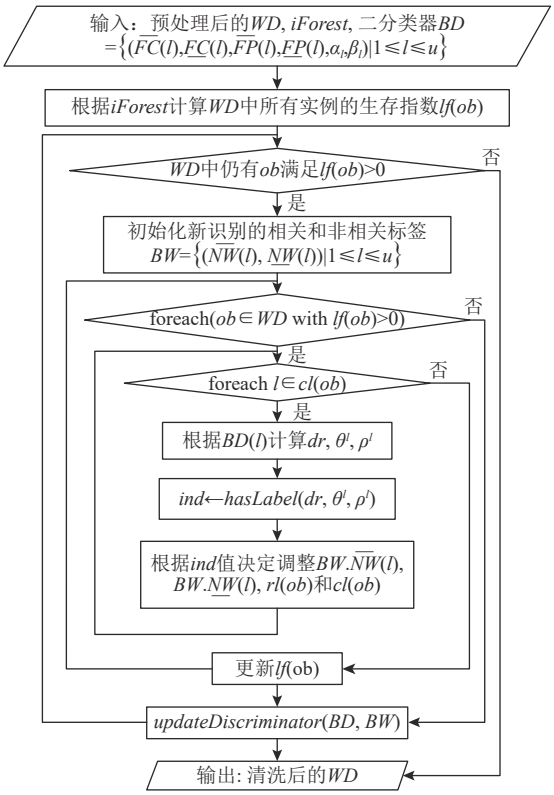

Fig. 3 Flow chart of iterative cleaning algorithm

图 3 迭代清洗算法流程图

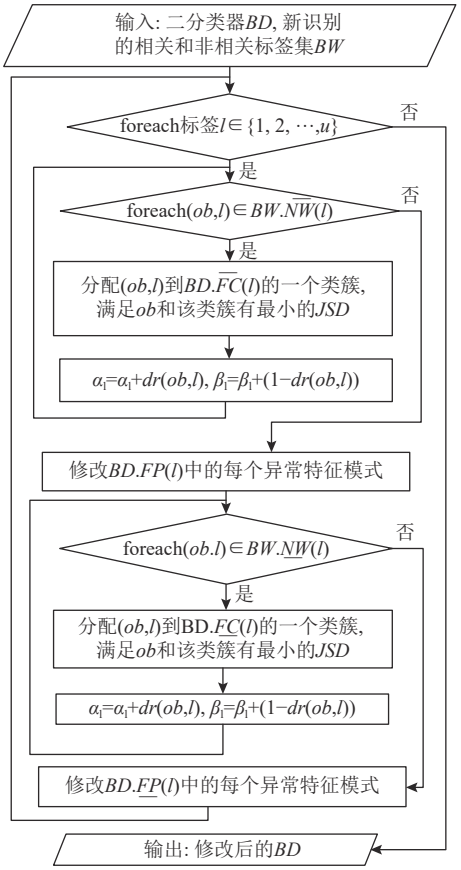

Fig. 4 Flow chart of *updateDiscriminator*

图 4 二分类器更新算法流程图

5 实验评估

实验在配备 AMD CPU（8 核@2.90 GHz）和 16 GB 内存的计算机上运行，原型系统用 Python 实现，

实验共采用了 3 个心电数据集，前 2 个是从社区医疗中心收集的真实数据集，每个样本是 12 导联、10 s 的记录，采样频率为 500 Hz. 异常标签共有 16 个，如表 1 所示. 一个数据集 CHE 包含 3 919 个样本，心电异常标签由专业医生标记和确认，标签是完整和正确的. 另一个数据集 CHW 包含 12 385 个样本，部分标签缺失或不正确. 第 3 个是 MIT-BIH 的公共数据集[45]，记为 MIT. MIT 收集了其中 40 个包含 II 和 VI 导联、30 min 的心电记录，取样频率是 360 Hz，将每个心电记录分成等长的 180 个长度是 10 s 的样本，将每个样本心跳对应的标签合并，作为该样本的多标签. 由于个别标签的样本非常稀疏，实验时采用了包含表 3 所示的 8 个异常标签的 7 166 个样本. 心电波去噪，基线漂移消除，QRS 波、P 波和 T 波的识别和特征提取按文献 [39] 所述实现，每个样本取 100 个特征.

Table 3 Abnormality Labels in MIT-BIH Dataset

表 3 MIT-BIH 数据集中的异常标签

标签名	标签名
左束支传导阻滞	右束支传导阻滞
心室融合心跳	房性期前收缩
交界性逸搏	正常心跳
室性期前收缩	异常房性早搏

为了度量标签清洗的效果，采用 3 个指标，即 *precision*, *recall*, *F*1，它们根据表 4 所示的 3 个指标定义.

Table 4 Meanings of *TP*, *FP* and *FN*

表 4 *TP*, *FP*, *FN* 的含义

指标名称	含义
TP	将正例预测为正例的数量
FP	将负例预测为正例的数量
FN	将正例预测为负例的数量

给定一个标签 l，其 *precision*, *recall*, *F*1 定义为

$$precision(l)=\frac{TP(l)}{TP(l)+FP(l)}, \tag{26}$$

$$recall(l)=\frac{TP(l)}{TP(l)+FN(l)}, \tag{27}$$

$$F1(l)=\frac{2\times precison(l)\times recall(l)}{precision(l)+recall(l)}. \tag{28}$$

汇报的度量根据标签的权重计算平均值. 例如，测试集 *TS* 上 *precision* 的计算为

$$precision=\sum_{l_i\in U}\frac{M_i^{TS}}{M^{TS}}\times PN(l_i), \tag{29}$$

其中 M_i^{TS} 为标签 l_i 在 *TS* 出现的次数，$M^{TS}=\sum_{i=1}^{u}M_i^{TS}$.

一方面，在真实的示例数据集 CHE 和弱数据集 CHW 按照如下步骤验证方法效果. CHW 作为弱数据集 *WD*，由于难以确定 *WD* 上的准确标签，根据训练的分类器效果间接度量标签的清洗效果. 首先，将 CHE 分为 2 部分：1/3 的 CHE 作为测试集 *TS*，其余作为示例数据集 *ED*，对 *WD* 的清洗效果按照 3 个步骤计算：

1）在 *WD* 上为每个异常训练 1 组二分类器. 然后，在 *TS* 上计算 *precision*, *recall*, *F*1，分别记为 $precision^{org}$, $racall^{org}$, $F1^{org}$.

2）在 *WD* 的清洗数据集上为每个异常训练 1 组二分类器，进而在 *TS* 上计算 *precision*, *recall*, *F*1，分别表示为 $precision^{cln}$, $recall^{cln}$, $F1^{cln}$.

3）上述 2 次测量值的差作为性能指标. 例如，对于 $df1=F1^{cln}-F1^{org}$ 作为性能指标.

另一方面,分别在 CHE 和 MIT 上模拟噪声标签,形成 2 个模拟数据集 SCHE 和 SMIT 来评估方法效果[7],即将各类标签按照一定的比率替换为不属于样本的随机标签,形成噪声标签. 具体地,从 CHE 中选择 1/3 的样本作为 *ED*,另外的 2/3 的样本生成 2 份拷贝. 一份作为正确标签参照,另一份引入不同级别(5%, 10%, 20%, 30%, 40%)的噪声标签作为 *WD*. 在 MIT 上也同样操作. 然后,对 *WD* 进行清洗,清洗后的样本与参照相对比,从而计算 *precision*, *recall*, *F*1.为避免实验结果的随机性,使用 6 折交叉验证计算各个度量. 根据采样效果,设置阈值 *st*=10.

5.1 影响标签发现规则和排除规则的因素

给定一个规则,其准确率 (*acc*) 是正确识别的正(或负)标签占识别出的正(或负)标签的比例. 下面分析在不同噪声水平下 2 种标签规则的影响因素. 除特别说明,本节汇报的是在 SCHE 上的结果,其他数据集上的结果呈现相同趋势,不再赘述.

5.1.1 影响标签发现规则的因素

图 5 和图 6 分别显示了在噪声水平为 10% 和 30% 时,固定其他参数,置信度阈值 *ct* 从 0.1 增加到 0.6 时,准确率 *acc* 的变化. 可以看出,随着 *ct* 的增加,准确率先增大,然后趋于平稳. 在其他噪声水平下,呈现类似的趋势. 这是因为 *ct* 越大,规则的置信度越高,规则的约束性更强,被包含的标签的准确度更高.

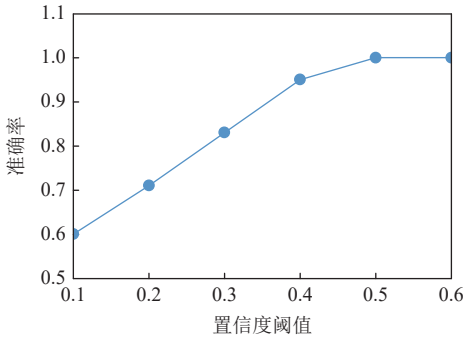

Fig. 5 *acc* changing with *ct* at noise level 10%

图 5 噪音水平 10% 时准确率随置信度阈值的变化

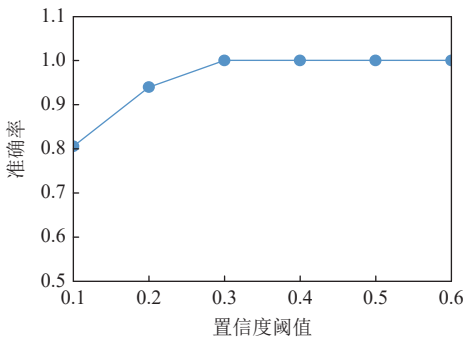

Fig. 6 *acc* changing with *ct* at noise level 30%

图 6 噪音水平 30% 时准确率随置信度阈值的变化

图 7 和图 8 分别显示了在噪声水平为 10% 和 30% 时,固定其他参数,正相关阈值 *rt* 从 0.1 增加到 0.6 时,准确率的变化. 可以看出,随着 *rt* 的增加,准确率先增大然后趋于平稳. 在其他噪声水平下,呈现类似的趋势. 这是因为正相关性越高,对 2 个标签共现频率的约束越高. 实验中,在模拟数据集上,对不同噪音水平采用不同的 *ct* 和 *rt*. 在真实数据集上,根据采样估计 *ct* 和 *rt*.

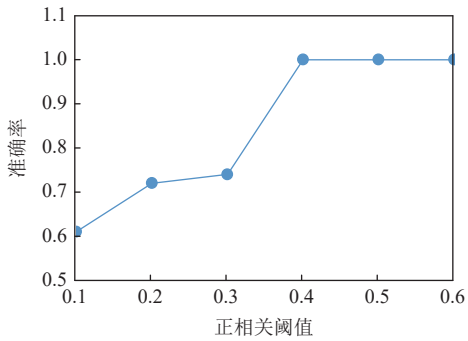

Fig. 7 *acc* changing with *rt* at noise level 10%

图 7 噪音水平 10% 时准确率随正相关阈值的变化

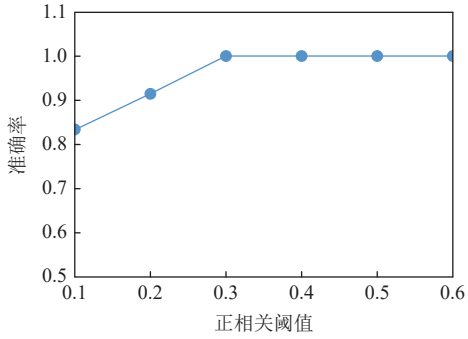

Fig. 8 *acc* changing with *rt* at noise level 30%

图 8 噪音水平 30% 时准确率随正相关阈值的变化

5.1.2 影响标签排除规则的因素

图 9 和图 10 分别显示了在噪音水平为 10% 和 30% 时,*acc* 随阈值 ε 的变化. 随着 ε 的增加,准确率先升高,然后降低. 这是因为,若 ε 太小,约束过于严格,会约束一些有效的标签排除规则,导致准确率受错误识别标签的影响;而随着 ε 变大,可以有效地发现更多排除规则,使得准确率趋于稳定;但 ε 进一步变大,也会导致排除规则准确率降低. 在其他噪音水平,呈现类似的效果.

5.2 消融实验

AFP 方法的标签清洗包含 3 个关键环节:第 1 步(ph1),在 *ED* 和 *WD* 上寻找共享异常特征模式,进而识别 *WD* 上的锚标签(是初始标签的一部分);第 2 步

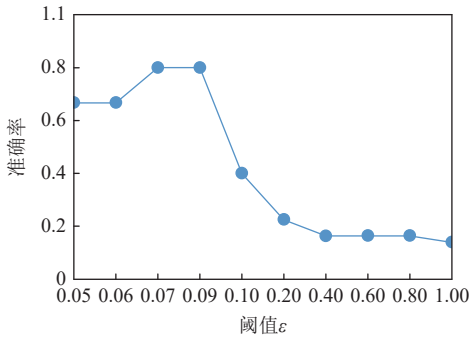

Fig. 9 *acc* changing with threshold ε at noise level 10%

图 9 噪音水平 10% 时准确率随阈值 ε 的变化

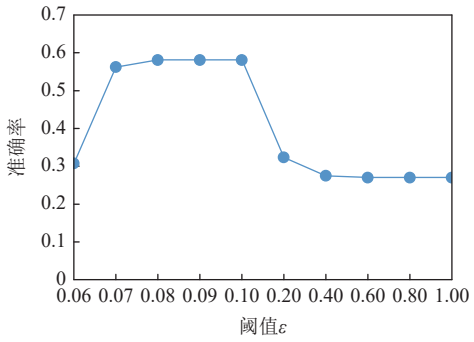

Fig. 10 *acc* changing with threshold ε at noise level 30%

图 10 噪音水平 30% 时准确率随阈值 ε 的变化

(ph2), 挖掘标签发现和排除规则, 然后扩充 *WD* 上样本的初始相关标签集; 第 3 步(ph3), 利用二分类器进行弱标签的迭代清洗. 为了验证各个环节的作用, AFP 方法分别消除 ph1, ph2, ph3, 记为 AFP-ph1, AFP-ph2, AFP-ph3 后, 汇报综合性能指标 *F*1 的变化情况.

图 11~15 汇报了噪声水平分别为 5%, 10%, 20%, 30%, 40% 时 2 个模拟数据集 SCHE 和 SMIT 上的消融实验结果. 图 16 汇报了在真实数据集 CHE 和 CHW 上的消融实验结果. 模拟和真实数据集上的结果表明:

1)在不同的噪声水平下去除步骤 ph1 后, 在模拟数据集 SCHE 上, *F*1 指标降低了 5.8~7.99 个百分点,

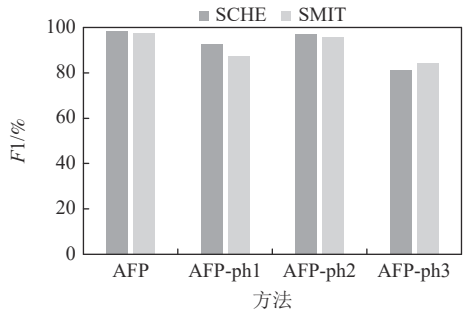

Fig. 11 Ablation experiment at noise level 5%

图 11 噪声 5% 的消融实验

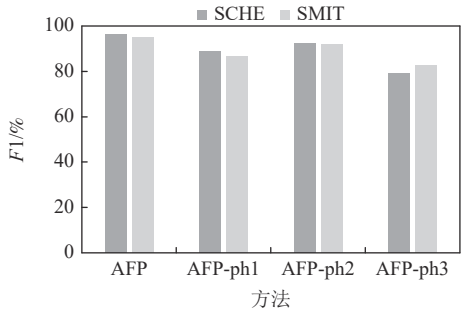

Fig. 12 Ablation experiment at noise level 10%

图 12 噪声 10% 的消融实验

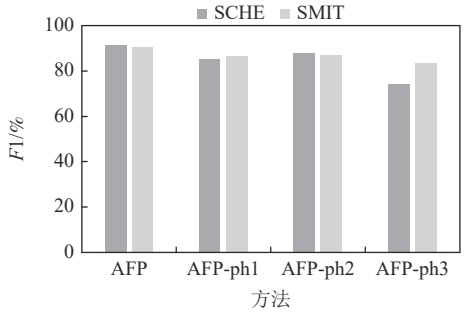

Fig. 13 Ablation experiment at noise level 20%

图 13 噪声 20% 的消融实验

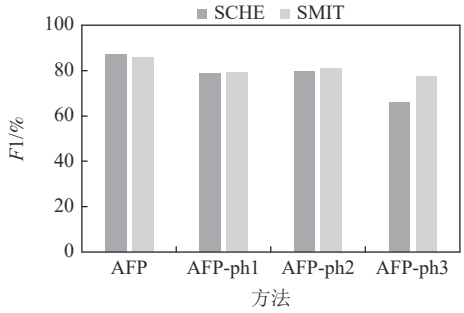

Fig. 14 Ablation experiment at noise level 30%

图 14 噪声 30% 的消融实验

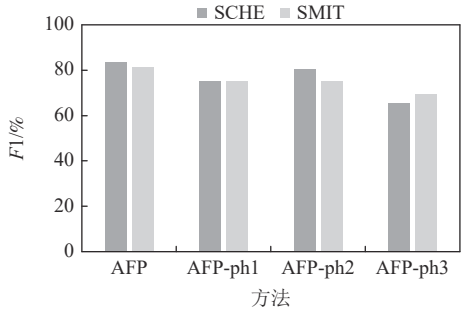

Fig. 15 Ablation experiment at noise level 40%

图 15 噪声 40% 的消融实验

SMIT 上降低了 6.21~10.17 个百分点, 在真实数据集上, *F*1 降低了 12.68 个百分点. 这是因为如果没有 ph1, 不仅不能确定 *WD* 的锚标签, 而且不能利用含锚标签的 *WD* 样本来扩充规则挖掘的可用样本.

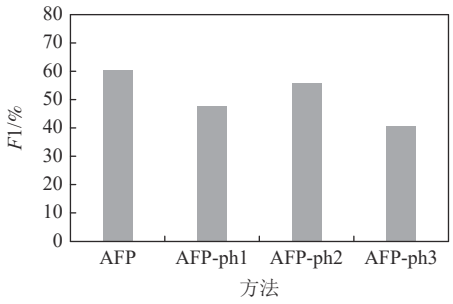

Fig. 16 Ablation experiment on real dataset

图 16 真实数据集上的消融实验

2) 在不同的噪声水平下去除步骤 ph2 后, SCHE 上的 *F*1 指标降低了 1.37~7.63 个百分点, SMIT 上降低了 1.75~6.12 个百分点, 真实数据集上降低了 4.43 个百分点. 这是因为, 步骤 ph2 根据挖掘的规则确定属于样本的异常标签, 在扩充初始相关标签集的同时, 尽量避免引入错误标签.

3)在不同的噪声水平下去除步骤 ph3 后, 模拟数据集 SCHE 上的 *F*1 指标降低了 16.89~20.93 个百分点, 在 SMIT 上 *F*1 降低了 7.07~13.25 个百分点, 真实数据集上的 *F*1 指标降低了 19.71 个百分点. 这是因为二分类器在清洗中不断自调整, 有效地识别单次清洗中无法识别的、处于分布边缘的标签. 可见, 步骤 ph3 居于 AFP 的主体地位.

5.3 比较研究

在模拟和真实数据集上, AFP 方法与交叉验证(cross validation, CV)方法[7] 和基于 DDF 的标签噪声过滤方法[38] 进行了比较. CV 方法利用 SVM, KNN, NB, LDA 和 DT 这 5 种分类器, 协同识别标记错误的样本. CV 方法为每个标签训练 5 个分类器, 如果 5 个分类器对一个实例的标签持不同认知, 则认为该标签被错误标记. 根据 3 个标准 S1, S2, S3 确定样本是否被错误标记. 对于 S1, 如果 5 个分类器都认为异常不属于实例, 则该异常是错误标签. 对于 S2, 如果 4 个或更多分类器认为异常不属于实例, 则认为该异常标记错误. 对于 S3, 如果 3 个或更多分类器认为异常不属于该实例, 则认为该异常标记错误. DDF 方法将每个样本的邻域样本划分为高密度和低密度区域, 然后针对不同的区域采用不同的噪声过滤规则进行过滤. 由于 DDF 能够识别出噪声标签, 但不能自动修补, 因此在模拟数据集上对每个标签计算 *precision* 时, 用 DDF 排除掉噪声标签后的该类标签数目作为识别出的该类标签数目.

图 17~21 汇报了 AFP, CV, DDF 方法在 SCHE 上的 *precision*, *recall*, *F*1 值. 可见, 当数据噪声级别为

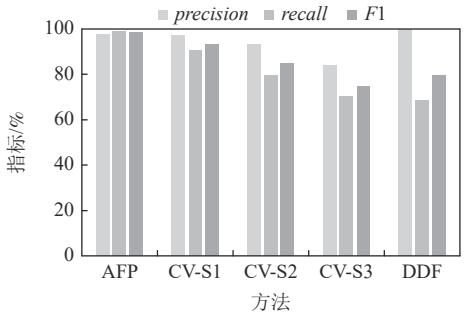

Fig. 17 Performance comparison over SCHE at noise level 5%

图 17 在 SCHE 上噪声 5% 时的性能比较

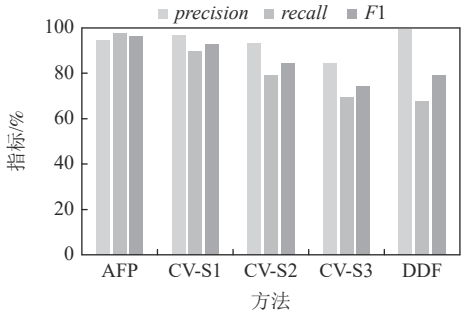

Fig. 18 Performance comparison over SCHE at noise level 10%

图 18 SCHE 上噪声 10% 时的性能对比

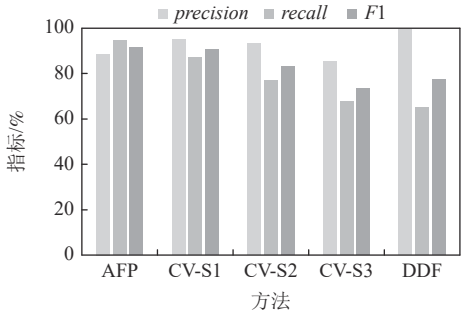

Fig. 19 Performance comparison over SCHE at noise level 20%

图 19 SCHE 上噪声 20% 时的性能对比

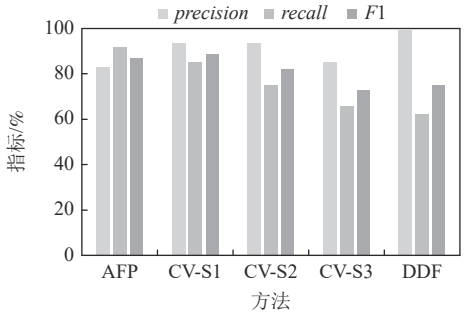

Fig. 20 Performance comparison over SCHE at noise level 30%

图 20 SCHE 上噪声 30% 时的性能对比

5% 时, AFP 方法的 *F*1 指标比 CV-S1 高 5.15 个百分点, 比 CV-S3 高 23.42 个百分点, 比 DDF 高 18.73 个

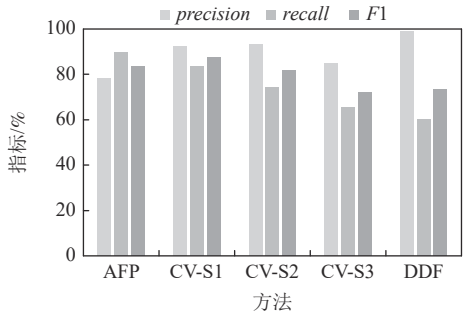

Fig. 21 Performance comparison over SCHE at noise level 40%

图 21 SCHE 上噪声 40% 时的性能对比

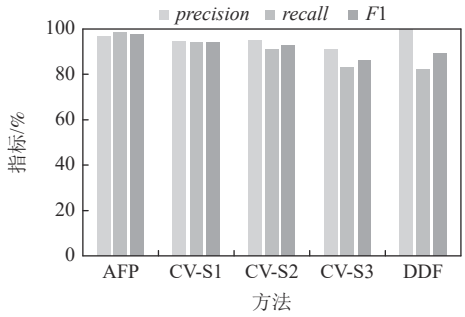

Fig. 22 Performance comparison over SMIT at noise level 5%

图 22 SMIT 上噪声 5% 时的性能对比

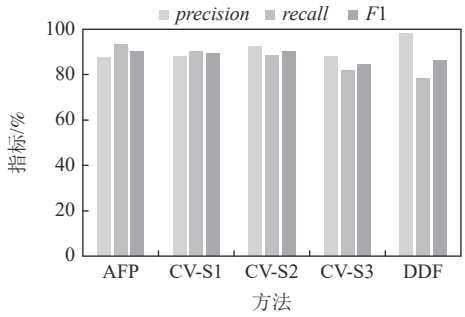

Fig. 23 Performance comparison over SMIT at noise level 20%

图 23 SMIT 上噪声 20% 时的性能对比

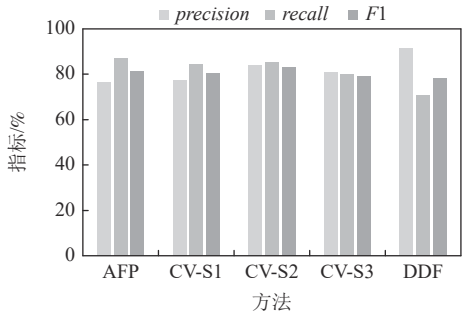

Fig. 24 Performance comparison over SMIT at noise level 40%

图 24 SMIT 上噪声 40% 时的性能对比

百分点. 当噪声水平为 10% 时, AFP 方法的 $F1$ 指标比 CV-S1 高 3.35 个百分点, 比 CV-S3 高 21.53 个百分点, 比 DDF 高 17.13 个百分点. 当噪声水平为 20% 时, AFP 方法的 $F1$ 指标比 CV-S1 高 0.7 个分点, 比 CV-S2 指标高 8.17 个百分点, 比 CV-S3 高 17.75 个百分点, 比 DDF 高 14.25 个百分点. 当噪声水平为 30% 时, AFP 方法的 $F1$ 指标比 CV-S1 低 1.63 个百分点, 但比 CV-S2 和 CV-S3 分别高 5.16 和 14.44 个百分点, 比 DDF 高 12.1 个百分点. 当噪声水平为 40% 时, AFP 方法的 $F1$ 指标比 CV-S1 低 3.93 个百分点, 比 CV-S2 和 CV-S3 分别高 1.86 和 11.18 个百分点, 比 DDF 高 9.84 个百分点. 实验结果表明, AFP 在 SCHE 的噪声不是很高的情况下, 清洗效果优于 CV 方法; 在数据噪声很高的情况下, AFP 方法略低于 CV-S1 方法, 仍优于 CV-S2 和 CV-S3; 同时, AFP 稳定地优于 DDF 方法.

图 22~24 汇报了 SMIT 上噪声为 5%, 20%, 40% 时的实验结果. 当噪声级别为 5% 时, AFP 方法的 $F1$ 指标比 CV-S1 高 3.25 个百分点, 比 CV-S2 高 4.6 个百分点, 比 CV-S3 高 11.33 个百分点, 比 DDF 高 8.16 个百分点. 当噪声水平为 20% 时, AFP 方法的 $F1$ 指标比 CV-S1 高 1.26 个分点, 比 CV-S2 高 0.11 个百分点, 比 CV-S3 高 6.09 个百分点, 比 DDF 高 4.28 个百分点. 当噪声水平为 40% 时, AFP 方法的 $F1$ 指标比 CV-S1 高 0.64 个百分点, 比 CV-S2 低 1.66 个百分点, 比 CV-S3 高 2.33 个百分点, 比 DDF 高 2.88 个百分点. 其他噪声水平下呈现类似趋势, 不再赘述. 实验结果表明, 在噪声不是很高的情况下, AFP 方法在 SMIT 上稳定地优于 CV 方法; 在噪声很高的情况下, AFP 方法仍优于 CV-S1 和 CV-S3 方法; 另外, AFP 稳定地优于 DDF 方法.

同时, 在真实数据集 CHE 和 CHW 上进行了比较. 首先, 在原始数据集 CHW 上训练分类模型, 然后分别用 AFP, CV, DDF 方法对数据集进行清洗, 比较在清洗前和清洗后的数据上训练分类器的性能指标. 表 5 显示了在真实数据集上的比较结果. AFP 方法的

Table 5 Comparison of AFP, CV and DDF on Real Dataset

表 5 真实数据集上 AFP，CV，DDF 方法的对比 %

方法	*precision*	*recall*	*F*1	*df*1
不清洗	57.74	55.95	55.24	
AFP	62.70	60.45	60.43	5.19
CV-S1	61.15	56.03	56.30	1.06
CV-S2	58.26	55.69	55.46	0.22
CV-S3	58.59	56.51	55.07	−0.17
DDF	63.65	57.41	58.37	3.13

平均 $F1$ 指标提高 5.19 个百分点, CV-S1 和 CV-S2 分别提高 1.06 和 0.22 个百分点, DDF 提高 3.13 个百分点. AFP 方法性能的优越主要因为 2 个原因: 首先, AFP 方法根据类簇在示例数据集和弱标签数据集上的一致性来识别锚异常, 充分利用了人工标注的知识, 也利用了弱标签数据集的可用信息. 其次, 采用迭代框架, 逐步缩小模糊区间来推断异常标签, 保证了清洗效果的可靠和稳步提升.

6 结 论

根据心电图(ECG)判断心脏异常是临床广泛应用的心脏健康检测技术. 目前, 自动异常检测主要采用有监督学习技术来实现. 由于生物电信号的多样性和相关性, 一个好的分类器通常需要依赖大量的高质量标签样本, 才能保证分类器的精度和泛化性. 这点对于当前流行的深度学习技术尤为重要. 然而, 高质量的心电标注不仅需要专业的心电知识, 而且要耗费大量的时间和精力. 实际中, 经常会有一些标注缺失或错误的心电数据集, 如何对这些弱标签的心电数据进行清洗, 提高标注质量, 使其变得可用, 是一个很有价值的问题.

设有一个包含所有正确标签的示例数据集, 可大可小, 这在实际中完全可行. 问题转化为在示例数据集的辅助下, 对弱标记数据集进行标签清洗, 将其转化为一个干净数据集. 由于一个心电异常通常表现出不同的特征模式, 提出了一种基于标签特征模式的标签清洗方法. 该方法首先确定高置信度属于实例的锚标签, 它们是相关标签集的子集. 然后, 以迭代方式清洗其他弱标签. 本文总结为 3 个方面: 1)根据示例数据和弱标签数据的一致性来识别锚特征模式, 充分结合了人工知识和数据的统计特性来提高标签区分能力. 2)提出了挖掘标签发现和排除规则的具体方法. 前者用于包含相关标签, 而后者用于删除无关标签. 3)采用迭代框架逐步清洗标签, 保证清洗效果的可靠和稳定. 在真实和模拟数据集上的实验结果证明了方法的有效性. 未来将研究根据病症的因果关系提高清洗效果.

作者贡献声明: 韩京宇负责论文思路、实验方案、论文撰写和修改; 陈伟和赵静负责实验和数据整理; 郎杭负责相关文献查阅和方法改进; 毛毅提供实验平台和专业知识指导.

参 考 文 献

[1] World Health Organization. Cardio-vascular diseases (CVDs)[EB/OL]. [2021-06-11]. https:// www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Liu Feifei, Liu Chengyu, Zhao Lina, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection[J]. Journal of Medical Imaging and Health Informatics, 2018, 8(7): 1368−1373

[3] Yang Hu. Diagnosis of myocardial infarction in electrocardiogram and recent progress [M]//Course Book of Electrocardiogram Specialty. Beijing: Beijing University Medical Press, 2005: 18−34 (in Chinese)
(杨虎. 心肌梗死心电图诊断与进展[M]//心电图专业人员培训教材. 北京: 北京大学医学出版社, 2005: 18−34)

[4] Tian Feng, Shen Xukun. Image annotation by semantic neighborhood learning from weakly labeled dataset[J]. Journal of Computer Research and Development, 2014, 51(8): 1821−1832 (in Chinese)
(田枫, 沈旭昆. 弱标签环境下基于语义邻域学习的图像标注[J]. 计算机研究与发展, 2014, 51(8): 1821−1832)

[5] Jin Linpeng, Dong Jun. Deep learning research on clinical electrocardiogram analysis[J]. SCIENTIA SINICA Informationis, 2015, 45(3): 398−416 (in Chinese)
(金林鹏, 董军. 面向临床心电图分析的深层学习算法研究[J]. 中国科学: 信息科学, 2015, 45(3): 398−416)

[6] Zheng Weizhe, Qiu Peng, Wei Juan. Sound recognition and detection based on multi-scale attention fusion in weak label environment[J]. Computer Science, 2020, 47(5): 120−123 (in Chinese)
(郑伟哲, 仇鹏, 韦娟. 弱标签环境下基于多尺度注意力融合的声音识别检测[J]. 计算机科学, 2020, 47(5): 120−123)

[7] Li Yaoguang, Cui Wei. Identifying the mislabeled training samples of ECG signals using machine learning[J]. Biomedical Signal Processing and Control, 2019, 47: 168−176

[8] Pasolli E, Melgani F. Genetic algorithm-based method for mitigating label noise issue in ECG signal classification[J]. Biomedical Signal Processing and Control, 2015, 19: 130−136

[9] Clifford G D, Liu Chengyu, Moody B, et al. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017[C/OL]//Proc of the 18th Computing in Cardiology(CinC). Piscataway, NJ: IEEE, 2017[2022-02-02]. https://cinc.org/archives/ 2017/pdf/065−469.pdf

[10] Cristina G V, Alexander B, Oriella G, et al. Two will do: Convolutional neural network with asymmetric loss, self-learning label correction, and hand-crafted features for imbalanced multi-label ECG data classification[C/OL]//Proc of the 22nd Computing in Cardiology. Piscataway, NJ: IEEE, 2021[2022-02-02]. https://www.cinc.org/archives/ 2021/pdf/CinC2021−024.pdf

[11] Frenay B, Verleysen M. Classification in the presence of label noise: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845−869

[12] Han Yufei, Sun Guolei, Shen Yun, et al. Multi-label learning with highly incomplete data via collaborative embedding[C]//Proc of the

24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1494−1503

[13] Wu Lei, Jin Rong, Jain A K. Tag completion for image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(3): 716−727

[14] Zhou Zhihua. A brief introduction to weakly supervised learning[J]. National Science Review, 2017, 5(1): 44−53

[15] Varma P, Ré C. Snuba: Automating weak supervision to label training data[J]. Proceedings of the VLDB Endowment, 2018, 12(3): 223−236

[16] Lee W S, Liu Bing. Learning with positive and unlabeled examples using weighted logistic regression[C]//Proc of the 20th Int Conf on Machine Learning. Palo Alto, CA: AAAI, 2003: 448−455

[17] Na B, Kim H, Song K, et al. Deep generative positive-unlabeled learning under selection bias[C]// Proc of the 29th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2020: 1155−1164

[18] Dong Haochen, Li Yufeng, Zhou Zhihua. Learning from semi-supervised weak-label data [C]// Proc of the 32nd AAAI on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 2926−2933

[19] Ding Jiaman, Liu Nan, Zhou Shujie, et al. Semi-supervised weak-label classification method by regularization[J]. Chinese Journal of Computers, 2022, 45(1): 69−81 (in Chinese)
(丁家满, 刘楠, 周蜀杰, 等. 基于正则化的半监督弱标签分类方法[J]. 计算机学报, 2022, 45(1): 69−81)

[20] Ding Hu, Xu Jinhui. Random gradient descent tree: A combinatorial approach for SVM with outliers [C]// Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2561−2567

[21] Xu Guibiao, Cao Zheng, Hu Baogang, et al. Robust support vector machines based on the rescaled hinge loss function[J]. Pattern Recognition, 2017, 63: 139−148

[22] He Fengxiang, Liu Tongliang, Geoffrey I W, et al. Instance-dependent PU learning by Bayesian optimal relabeling [J]. arXiv preprint, arXiv: 1808. 02180, 2018

[23] Basile T M A, Mauro N D, Esposito F, et al. Density estimators for positive-unlabeled learning[M]// New Frontiers in Mining Complex Patterns. Berlin: Springer, 2017: 49−64

[24] Chaudhari S, Shevade S. Learning from positive and unlabelled examples using maximum margin clustering[C]// LNCS 7665: Proc of the 19th Int Conf on Neural Information Processing. Berlin: Springer, 2012: 465−473

[25] Gong Chen, Shi Hong, Liu Tongliang, et al. Loss decomposition and centroid estimation for positive and unlabeled learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 918−932

[26] Zhang Minling, Zhou Zhihua. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819−1837

[27] Gibaja E, Ventura S. A tutorial on multilabel learning[J]. ACM Computing Surveys, 2015, 47(3): 1−38

[28] Boutell M R, Luo Jiebo, Shen Xipeng, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757−1771

[29] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3): 333−359

[30] Fürnkranz J, Hüllermeier E, Mencía E L, et al. Multilabel classification via calibrated label ranking[J]. Machine Learning, 2008, 73(2): 133−153

[31] Tsoumakas G, Katakis I, Vlahavas I. Random K-Labelsets for multi-label classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1079−1089

[32] Zhang Minling, Zhou Zhihua. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038−2048

[33] Clare A, King R D. Knowledge discovery in multi-label phenotype data[C]//Proc of the 5th European Conf on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 2001: 42−53

[34] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C]// Proc of the 14th Int Conf on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA: MIT Press, 2001: 681−687

[35] Li Feng, Miao Duoqian, Zhang Zhifei, et al. Mutual information based granular feature weighted k-nearest neighbors algorithm for multi-label learning[J]. Journal of Computer Research and Development, 2017, 54(5): 1024−1035 (in Chinese)
(李峰, 苗夺谦, 张志飞, 等. 基于互信息的粒化特征加权多标签学习K近邻算法[J]. 计算机研究与发展, 2017, 54(5): 1024−1035)

[36] Liu Tongliang, Tao Dacheng. Classification with noisy labels by importance reweighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 447−461

[37] Himanshu K, Naresh M, Sastry P S. Robust learning of multi-label classifiers under label noise[C]// Proc of the 7th ACM India Special Interest Group on Knowledge Discovery and Data Mining. New York: ACM, 2020: 90−97

[38] Chen Qingqiang, Wang Wenjian, Jiang Gaoxia. Label noise filtering based on the data distribution[J]. Journal of Tsinghua University: Science and Technology, 2019, 59(4): 262−269 (in Chinese)
(陈庆强, 王文剑, 姜高霞. 基于数据分布的标签噪声过滤[J]. 清华大学学报: 自然科学版, 2019, 59(4): 262−269)

[39] Han Jingyu, Sun Guangpeng, Song Xinhai, et al. Detecting ECG abnormalities using an ensemble framework enhanced by Bayesian belief network[J]. Biomedical Signal Processing and Control, 2022, 72(A): 103320

[40] Liu F T, Ting K M, Zhou Zhihua. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1−39

[41] Ferguson T S. A Bayesian analysis of some nonparametric problems[J]. The Annals of Statistics, 1973, 1(2): 209−230

[42] David M B, Michael I J. Variational methods for the Dirichlet process[C]// Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 89−96

[43] Černý V. Thermo dynamical approach to the traveling salesman problem: An efficient simulation algorithm[J]. Journal of Optimization Theory and Applications, 1985, 45: 41−51

[44] Han Jiawei, Kamber M, Pei Jian. Data Mining: Concepts and

Techniques[M]. 3rd ed. San Francisco: Morgan Kaufmann, 2012: 38-47

[45] George M, Roger M. MIT-BIH Arrhythmia Database [DB/OL]. (2005-02-24)[2021-03-07]. https://physionet.org/content/mitdb/1.0.0/

Han Jingyu, born in 1976. PhD, professor. Member of CCF. His main research interests include biomedical information processing, database system, and machine learning.

韩京宇, 1976 年生. 博士，教授. CCF 会员. 主要研究方向为生物信息处理、数据库系统和机器学习.

Chen Wei, born in 1995. Master candidate. His main research interests include biomedical information processing and machine learning.

陈　伟, 1995 年生. 硕士研究生. 主要研究方向为生物信息处理和机器学习.

Zhao Jing, born in 1996. Master. Her main research interests include machine learning and database systems.

赵　静, 1996 年生. 硕士. 主要研究方向为机器学习和数据库系统.

Lang Hang, born in 1999. Master candidate. His main research interests include machine learning and bioinformatics.

郎　杭, 1999 年生. 硕士研究生. 主要研究方向为机器学习和生物信息学.

Mao Yi, born in 1985. PhD, lecturer. Her main research interests include biomedical information processing and machine learning.

毛　毅, 1985 年生. 博士，讲师. 主要研究方向为生物信息处理和机器学习.