第52卷第5期
2020年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 52 No. 5
Oct. 2020

DOI:10.16356/j.1005-2615.2020.05.007

非结构化数据的多粒度集成分类方法

王子一[1,2]，徐苏平[1,2]，商　琳[1,2]

（1.计算机软件新技术国家重点实验室(南京大学)，南京，210023；2.南京大学计算机科学与技术系，南京，210023）

摘要：深度学习模型已经在文本和图像等分类任务上取得了不错的效果，然而深度学习模型很难为分类结果提供可解释性。本文提出一种非结构化数据的多粒度集成分类方法，与其他学习方法相比，多粒度集成分类方法能够保留数据的上下文信息。在多粒度集成分类方法中，数据被划分成不同的粒度，用于训练不同的基学习器，这些学习结果为集成模型最后的分类提供了可解释性。基学习器根据它们在验证集上的精度被赋予不同的权重，从而构造出一个较好的集成学习器。在实验中，本文验证了所提出模型在3种非结构化数据类型（文本、医学图像和时间序列）上的有效性。实验结果表明，本文的模型比现有的基准方法简单，具有较好的分类精度，并且能够为数据的分类提供可解释性。

关键词：集成学习；多粒度；神经网络；分类

中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1005-2615(2020)05-0723-06

Multi-grained Ensemble Classification Method for Unstructured Data

WANG Ziyi[1,2]，*XU Suping*[1,2]，*SHANG Lin*[1,2]

（1. State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing, 210023, China; 2. Department of Computer Science and Technology, Nanjing University, Nanjing, 210023, China）

Abstract: Deep learning model has achieved great success in text and image classification tasks. However, it is difficult for a deep learning model to provide an interpretable classification intuition. In this paper, we propose a multi-grained ensemble classification method for unstructured data. Compared with other learning methods, our multi-grained ensemble classification method can preserve the context information of data. In the multi-grained ensemble classification method, data are divided into different granularity, and the data with different granularity are used to train different base learners. Their learning results provide interpretability for the final classification of the ensemble model. Base learners are assigned by different weights according to their performances on the validation set, therefore, a better ensemble learner can be constructed. In the experiments, we verify the validity of our model on three types of unstructured data (i.e., text, medical images, and time series), and experimental results show that our model is not only simpler than the state of art but also competitive in the accuracy. Meanwhile, it can provide the interpretability for the final classification results.

Key words: ensemble learning; multi-granularity; neural networks; classification

集成方法通过同时训练多个学习器然后将它们结合起来完成学习任务，已经在很多现实任务上取得了巨大的成功。事实上，集成学习几乎在所有学习任务上都非常有用[1]。计算机视觉、自然语言处理和时间序列分析等都从集成方法中受益匪浅。Zhou等[2]提出了一种两层的集成结构模型用于肺

基金项目：国家自然科学基金(61672276，51975294)资助项目。

收稿日期：2019-09-01；**修订日期：**2020-02-10

通信作者：商琳，女，教授，E-mail：shanglin@nju.edu.cn。

引用格式：王子一，徐苏平，商琳．非结构化数据的多粒度集成分类方法[J]．南京航空航天大学学报，2020，52(5)：723-728. WANG Ziyi, XU Suping, SHANG Lin. Multi-grained ensemble classification method for unstructured data[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 723-728.

癌细胞的识别。在阿尔茨海默症的早期诊断中,早期方法通常只考虑脑电图中的单通道数据,为了能够同时利用多个通道的数据,Polikar等[3]提出了一种集成学习方法,该方法利用不同电极上采集的不同数据训练不同的基学习器,以响应不同的刺激和不同的频带,并将它们的输出结合起来作为最终的诊断。Bagnall等[4]提出了一种简单的基于变换和集成的时间序列分类方法,比绝大多数时间序列分类方法有更高的精度。Bag-of-SFA-symbols (BOSS)[5]模型使用一种符号化的表示方法将时间序列子序列的提取和对噪声的鲁棒性结合起来,然后集成多个学习器完成最后的分类任务,BOSS在时间序列分类上取得了极大的成功。文本情感分类是自然语言处理中一个热门的研究领域,在绝大多数的文本情感分类比赛中,单一模型很难达到最好的性能,集成多个分类模型能够提升文本情感分类的精度。以上提到的这些集成方法都在特定的任务上取得了非常好的性能,然而除了分类的准确性之外,通常还希望对数据有一定的了解,希望模型能够为分类结果提供可解释性。

现实中的数据往往存在着冗余和噪声,因此对大多数分类任务而言,数据的不同特征对于分类有着不同的重要性。一个具有可解释性的分类模型通常能够在分类学习时,对人们应该更加关注数据的特征加以指导。对于处理结构化数据的分类任务而言,有很多特征选择的方法为分类提供了可解释性。然而对于非结构化数据,分类模型很难为非结构化数据提供可解释性。对非结构化数据进行分类时往往需要先对原始数据进行高级特征提取,比如对图像边缘信息进行提取、对文本进行嵌入、对时间序列进行傅里叶变换和小波变换等,然后对提取好的特征进行特征选择或降维等处理,最后进行分类。虽然可以对非结构化数据进行高级特征提取后的特征进行特征选择,然而需要理解这些特征仍然很难,比如单词的嵌入向量,对时间序列进行傅里叶变换后的傅里叶系数等,因此很难理解非结构化数据的分类结果。目前,深度学习模型在处理非结构化数据的分类任务上取得了令人瞩目的成绩,然而深度学习模型本身是一种黑箱模型,因此理解深度学习模型处理非结构化数据时的分类可解释性更加困难。

虽然处理非结构化数据的分类模型为分类提供可解释性很难,但是仍有一些工作在这方面进行了探索。(Class activation map (CAM)[6]是卷积神经网络(Convolutional neural networks, CNN)中的一种可视化方法,为CNN分类提供了可解释性。CAM是利用CNN中的全局平均池化层生成的,它可以精准地突出图像中某些区域应用于分类学习。另一种CAM的改进算法,带有梯度加权的CAM (Grad-CAM)[7]使用了任何目标概念的梯度,流入最后一层卷积层,生成一个粗定位的映射,这样能突出图像中预测的重要区域Shapelet[8]是一种经典的时间序列分类方法。Shapelet是时间序列中的一个子段,某种程度上能够最大程度地表示一类时间序列的关键特征。Shapelet除了能够提高时间序列的分类精度,还能为时间序列分类提供可解释性,Shapelet在某种意义上表示时间序列在分类时更应该关注的区域。

如上所述,虽然目前有一些方法在处理非结构化数据分类的同时还能为分类提供可解释性,但是这些方法很难同时兼顾分类精度和可解释性。在CAM网络中,使用全局平均池化来代替全连接,这样可以保留原始图像的位置信息,并通过对最后一层卷积层后的Feature map进行加权就可以对分类的关键性区域进行可视化。使用全局平均池化代替全连接虽然可以保留原始图像中的位置信息,但是使用CAM网络分类会大幅度损失网络的分类精度。在时间序列中使用Shapelet用于分类,使得分类模型有很好的可解释性,但是单一的Shapelet子段很难有效地利用整个时间序列的所有特征。目前,虽然有方法能够同时学习时间序列的多个Shapelet,但是这些Shapelet仍不能包含时间序列中全部的有效信息。

为了能够同时兼顾模型在非结构化数据上的分类精度和尽可能为分类提供可解释性,本文提出了非结构化数据的多粒度集成分类方法。在非结构化数据的多粒度集成分类方法中,本文首先对原始数据进行多粒度划分。不同粒度的数据被用来训练不同的基学习器,与其他的学习方法不同的是,非结构化数据的多粒度集成分类方法能够保留数据的上下文信息。不同的基学习器表示原始数据中不同区域的学习结果,这为最终的集成分类模型提供了可解释性。在集成分类阶段,不同基学习器根据其在验证集上的精度被赋予不同的权重,通过对每个基学习器输出结果的加权,最后得到集成分类的结果。

1 非结构化数据的多粒度集成分类方法

本节将详细介绍所提出的非结构化数据的多粒度集成分类方法。首先给出非结构化数据多粒度划分的定义,并讨论如何利用不同的基学习器的学习结果对最后的分类结果提供可解释性;接下来

介绍如何选择合适的分类器作为基学习器以及如何计算基学习器的权重；最后给出了4种多样性度量指标来度量不同粒度数据训练出的基学习器之间的多样性。非结构化数据的多粒度集成分类方法的框架如图1所示，整个框架通过迭代的方式来确定基学习器的数量。

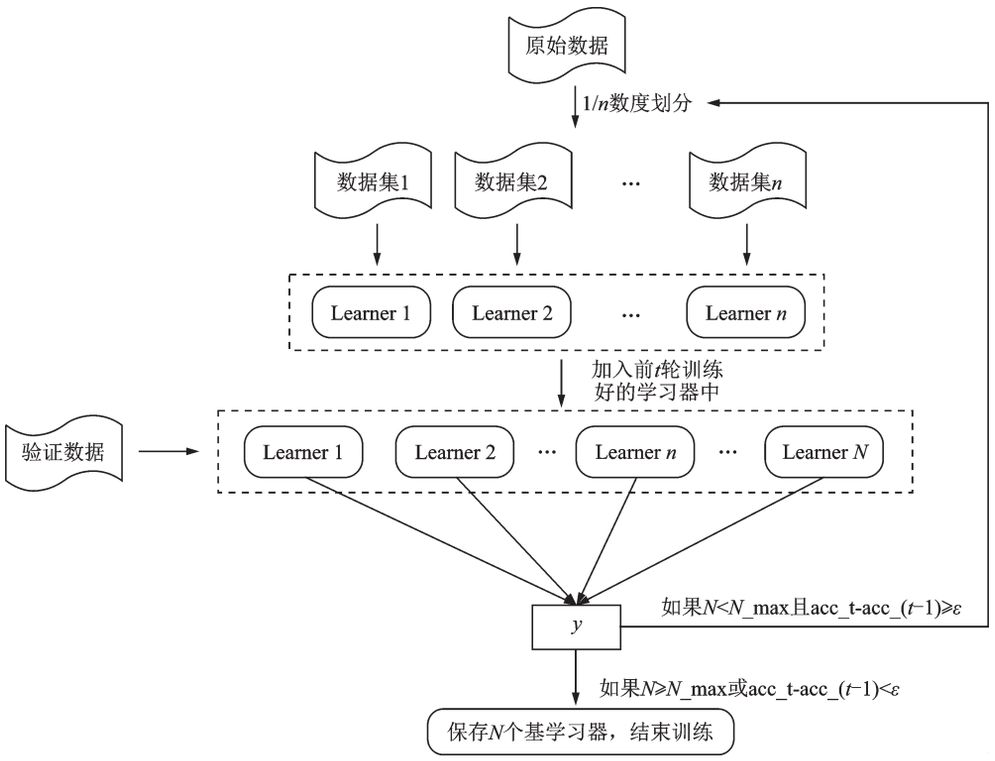

图1　非结构化数据的多粒度集成分类方法框架

Fig.1　Framework of the multi-grained ensemble classification method for unstructured data

1.1　多粒度划分

对数据样本进行扰动是集成学习里提高集成多样性中最常见的方法。最常见的样本扰动方法有自助采样法和序列采样法[1]，随机森林使用的是自助采样法，AdaBoost使用的是序列采样法。与这两种方法不同的是，在非结构化数据的多粒度集成分类方法中，本文通过对样本进行不同粒度的划分来进行样本扰动。

$1/n$粒度表示对数据的每个维度进行n等分。1个长度为L的一维数据，比如图像，对其进行$1/n$粒度划分后会得到n^2个大小为$\mathrm{Floor}(W/n)\times\mathrm{Floor}(H/n)$的样本，如图2所示。

对数据进行不同粒度划分，可以得到不同的数据集，与自助采样和序列化采样不同，通过多粒度划分得到的数据集中没有完全相同的样本。对数据进行不同粒度的划分不仅可以提高集成的多样性，还可以保留数据的上下文信息。不同基学习器都代表着原始数据中不同区域的学习结果，通过训练不同的基学习器，观察这些基学习器在验证集上的表现可以判断原始数据中哪部分区域对于分类来说是重要的。一般来说，基学习器在验证集上的精度越高，该基学习器所对应的区域对于分类来说越重要。

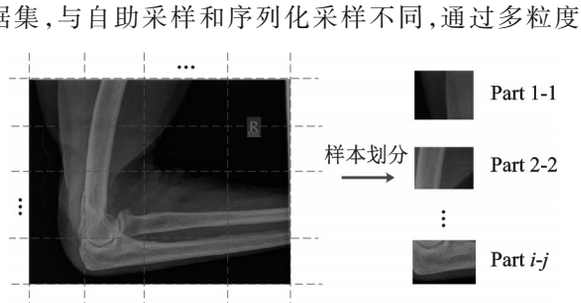

图2　二维图像的$1/n$粒度划分

Fig.2　$1/n$ granularity divide of a 2-D image

现实中收集的数据往往存在很多冗余，数据中有很多噪声，对分类有用的区域可能只占原始数据的一部分。对原始数据进行多粒度划分可以让不同的基学习器关注数据的不同区域，降低学习难度，同时通过集成多个学习器的结果可以提高最终的分类性能。

1.2　基学习器选择

集成的多样性一般指基学习器之间的差异性，是集成学习中最关键的问题。为了构造一个好的集成学习器关键在于所有的基学习器应该“好而不同”[1]。一般而言，“好”的学习器是指学习器性能好（精度高或均方误差小）。学习器之间的“不同”很难定义，一般来说有4种策略来增强学习器之间

的多样性:数据样本扰动、输入属性扰动、输出表示扰动,以及算法参数扰动[1]。在集成学习中,不同的多样性增强策略,应该使用不同的基学习器。在多粒度集成分类方法中,使用样本扰动策略来提高集成多样性,因此应该选择对样本扰动敏感的分类器作为基学习器。一般而言,学习器可以被分为稳定的基学习器和不稳定的基学习器。稳定的基学习器是指对样本扰动不敏感的基学习器,比如,支持向量机、朴素贝叶斯和k近邻学习器等。不稳定的基学习器是指对样本扰动敏感的学习器,样本的轻微扰动会导致学习器学习性能的显著变化,如:决策树和神经网络。因此,应该选择决策树或者神经网络作为基学习器,但是对非结构化数据,神经网络有着比决策树更好的分类效果。本文选择神经网络作为基学习器。

在集成学习中基学习器数量决定了最后集成学习器的性能。一般而言,集成学习器的性能随着基学习器数量的增加而提高,但并不是严格意义上的单调递增。考虑实际应用中基学习器的训练代价和存储代价,因此需要选择合适数量的基学习器来构造集成学习器。目前,有两种主流的确定基学习器数量的方法:一种是静态的,另一种是动态的。静态的方法一般是直接确定一个最大的基学习器数量$N_\max$,然后学出$N_\max$个基学习器为止。动态方法如图1所示,第t次迭代对数据进行$1/n$粒度的划分,得到n个不同的数据集用于训练n个不同的基学习器。根据当前所有基学习器在验证集上精度acc_t与第$t-1$次中所有基学习器在验证集上的精度$\mathrm{acc}_(t-1)$进行比较,如果当前学习器数量$N < N_\max$且$\mathrm{acc}_t-\mathrm{acc}_(t-1) \geqslant \varepsilon$,那么对数据进行更细粒度的划分,否则停止划分。

1.3 基学习器权重计算

在集成学习中,不同的集成学习算法有不同的基学习器结合方式。在Bagging中,所有基学习器被设置为相等的权重。在Boosting(如AdaBoost)中,每个基学习器根据其在训练集上的错误率被赋予不同的权重。在多粒度集成分类方法中,不同的基学习器对应原始数据中不同区域的学习结果。由于数据存在冗余,因此数据不同部分对于分类的重要性不一样,在结合多个基学习器时,应该对不同的基学习器赋予不同的权重。根据基学习器在验证集上的精度,赋予基学习器不同的权重。若一共有N个基学习器被用来构造集成学习器,且这N个基学习器在验证集上的精度都大于随机猜测的精度,那么可以根据它们在验证集上的精度得到N个不同的基学习器权重(W_1, W_2, …, W_N)。在本文提出的方法中,对基学习器进行加权投票时并不直接使用N个基学习器的权重。在最终加权投票时设定了一个投票权重的置信度值$C_i=W_i-1/c$,其中,c表示数据集中类别数,$1/c$表示类别平衡下的随机猜测的精度,在多粒度集成分类方法中使用C_i作为第i个基分类器最后投票的权重,这样使得最后的集成学习器更加偏好强基学习器的学习结果。

在测试过程中,所有的样本同样按照训练阶段的粒度划分方式进行划分。将划分好的数据输入到训练好的基学习器中,最后将每个基学习器的输出进行加权得到最终的分类结果。

2 实验和结果分析

本节使用提出的非结构化数据的多粒度集成分类方法在3种不同的非结构化数据(图像、文本和时间序列)上进行实验。

2.1 医学图像分类实验

2.1.1 数据集

本节实验的数据集来自目前最大的Andrew Ng公开的医学影像数据集MURA[9],该数据集包含了人体7个部位的X光片,分别对应7个不同的数据集,从中选取了5个数据集进行了实验。数据集信息如表1所示,该数据集中随机地选取70%的数据用于训练,30%的数据用于测试。

表1 MURA数据集的统计信息

Table 1 Statistics of MURA datasets

Study	Finger	Hand	Humerus	Shoulder	Wrist
Normal	1 389	1 613	411	1 479	2 295
Abnormal	753	602	367	1 594	1 451
Total	2 142	2 215	778	3 073	3 746

2.1.2 实验结果和分析

模型训练过程中使用Resnet作为基学习器,使用ImageNet预训练好的参数初始化Resnet。在粒度划分阶段使用静态的方式选定基学习器的数量,规定最大的基学习器数量为$N(N=1+4+9=14)$。最终14个分类器的分类结果如表2所示。

从表2中可以看出,不同基学习器在测试集上的精度有着显著的区别,这意味着原始数据中的不同区域训练出的基学习器对于分类有着不同的重要性。以Finger数据集为例,第10个基学习器的性能最好,第9个基学习器的性能最差。因此,在原始数据集中可能第10个基学习器所对应的区域对于分类来说是最重要的,而第9个基学习器所对应的区域可能没有那么重要,基学习器的分类结果为图像的分类提供了可解释性。最后,所有基学习器集成后的学习器在所有数据集上都取得了最好的分类效果,这说明集成能够提高分类精度。

表 2　15个学习器在5个数据集上的测试精度

Table 2　Testing accuracy of 15 learners on five datasets

Leaner	Finger	Hand	Humerus	Shoulder	Wrist
1	0.757	0.750	0.840	0.755	0.839
2	0.737	0.734	0.773	0.773	0.778
3	0.761	0.726	0.767	0.739	0.797
4	0.735	0.715	0.750	0.723	0.797
5	0.718	0.741	0.732	0.725	0.783
6	0.729	0.684	0.736	0.685	0.721
7	0.729	0.713	0.809	0.689	0.736
8	0.705	0.702	0.687	0.698	0.740
9	0.692	0.728	0.732	0.719	0.748
10	0.764	0.730	0.795	0.751	0.827
11	0.716	0.691	0.712	0.700	0.734
12	0.707	0.663	0.705	0.698	0.734
13	0.729	0.702	0.729	0.694	0.786
14	0.703	0.676	0.753	0.682	0.713
Our model	**0.774**	**0.772**	**0.844**	**0.800**	**0.850**

2.2　文本情感分类实验

2.2.1　数据集

IMDB[10]是文本情感分类中常见的数据集,包含25 000条电影的评论用于训练,25 000条电影评论用于测试。在IMDB中对电影的评分范围是[1,10],在实验时将其线性映射到[0,1]。

2.2.2　实验结果和分析

实验使用CNN[11]作为基学习器,并且使用预训练好的google word2vec的参数(300维)初始化词向量。在粒度划分阶段使用静态方式选定基学习器的数量,规定最大的基学习器数量为N(N=1+2+3+4=10),使用10个基学习器构造出最后的集成学习器。实验部分将本文提出的算法与SVM (Support vector machine),CNN,MLP(Multi-layer perceptron)和RF (Random forest)[12]进行了对比分析,算法在测试集上的分类精度如表3所示,10个基学习器在测试集上的性能如图3所示。

表 3　IMDB数据集上测试精度比较

Table 3　Comparison of testing accuracy on IMDB

分类算法	CNN	SVM	MLP	RF	Our model
测试精度	0.880	0.876	0.880	0.853	**0.887**

从表3中可以看出,在IMDB数据集上本文的方法取得最高的分类精度。从图3中可以看出,对文本进行粒度划分会降低文本的分类精度,电影评论的后半部分可能包含更多的情感信息。

图3是在IMDB采用了最多10个不同基学习器集成后的分类结果,横坐标表示采用的基学习器数量,纵坐标对应分类结果。由图可见,Learner 3性能大于0.82,好于Learner 2(低于0.82),Learner 6

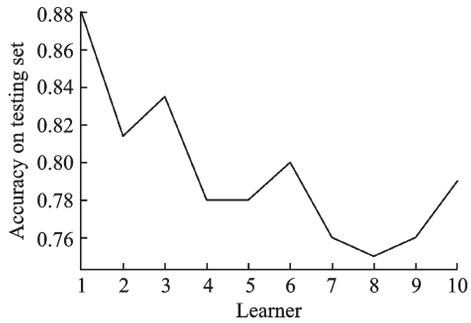

图3　10个不同的基学习器在IMDB上的测试精度

Fig.3　Testing accuracy of 10 learners on IMDB

性能(0.80)好于Learner 4(0.78)和Learner 5(0.78)。但对文本进行更细粒度划分会降低文本的分类精度(Learner 4的性能大大低于Learner 2),这说明电影评论的后半部分可能包含更多的情感信息,从而影响了分类结果。

2.3　时序分类实验

2.3.1　数据集

UCR数据集是目前时间序列分类中最常用的数据集,从UCR数据集中随机选取5个数据集进行实验,这5个数据集的信息如表4所示。

表 4　时间序列数据集的统计信息

Table 4　Statistics of time series datasets

Dataset	Train	Test	Classes	Length
Adaic	390	391	37	176
Gun_point	50	150	2	150
Haptics	155	308	5	1 092
ItalyPower	67	1 029	2	24
MedicalIma	381	760	10	99

2.3.2　实验结果和分析

本节实验使用带有一层隐层的MLP作为基学习器。在数据粒度划分阶段使用动态的方式决定基学习器的数量。实验部分与现有的3种常见的时间序列分类算法进行了比较,测试集上的误差如表5所示。

表 5　5个UCR时间序列数据集的测试误差

Table 5　Testing error for five UCR time series datasets

Dataset	DTW	BOSS	COTE	Our model
Adaic	0.396	0.220	0.233	**0.202**
Gun_point	0.093	**0.000**	0.007	0.007
Haptics	0.623	0.536	0.488	**0.474**
ItalyPower	0.050	0.053	0.036	**0.026**
MedicalIma	0.263	0.288	0.258	**0.221**

为了得到可解释性的分类结果,本文给出了3个基学习器在Gun_point测试集数据上的分类性能,分别为:Learner 1 = 0.933,Learner 2 = 0.887,

Learner 3 = 0.96。

从表5中可以看出,本文提出的方法在UCR数据集上非常有竞争性,在4个数据集上都取得了最好的效果,其中BOSS[5]与COTE[4]都是针对时间序列设计的集成学习方法。从实验结果中给出的3个基学习器在Gun_point测试集数据上的分类性能可以看出,Learner 3的分类性能要显著优于Learner 2的分类性能,这说明在Gun_point数据集上后半段时间序列包含了更多对于分类有用的信息。参考对Gun_point数据集进行Shapelet的搜索结果,Gun_point数据集中,Shapelet出现在数据的后半部分,从而证实了本文的方法在分类上除了精度高的优势外,还具有一定的可解释性特点。

3 结 论

本文提出了一种非结构化数据的多粒度集成分类方法。通过对数据进行多粒度划分,可以提高集成的多样性并且能够保留数据的上下文信息。不同的基学习器对应原始数据中不同区域的学习结果,为最后的分类结果提供了可解释性。在实验部分,本文在3种不同类型的非结构化数据上进行效果验证,实验结果表明,本文的方法在这3种非结构化的数据集上均取得了非常好的效果。

参考文献:

[1] ZHOU Z H. Ensemble methods: Foundations and algorithms[M]. Boca Raton, FL, USA: CRC Press, 2012.

[2] ZHOU Z H, JIANG Y, YANG Y B, et al. Lung cancer cell identification based on artificial neural network ensembles[J].Artificial Intelligence in Medicine, 2002, 24(1): 25-36.

[3] POLIKAR R, TOPALIS A, PARIKH D, et al. An ensemble based data fusion approach for early diagnosis of Alzheimer's disease[J]. Information Fusion, 2008, 9(1): 83-95.

[4] BAGNALL A, LINES J, HILLS J, et al. Time-series classification with COTE: The collective of transformation-based ensembles[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(9): 2522-2535.

[5] SCHAEFER P. The BOSS is concerned with time series classification in the presence of noise[J]. Data Mining and Knowledge Discovery, 2015, 29(6): 1505-1530.

[6] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2921-2929.

[7] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Venice Italy: IEEE, 2016: 618-626.

[8] YE L X, KEOGH E. Time series shapelets: A new primitive for data mining[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2009: 947-956.

[9] RAJPURKAR P, IRVIN J, BAGUL A, et al. MURA: Large dataset for abnormality detection in musculoskeletal radiographs[EB/OL].[2019-07-10]. arXiv: 1712.06957, 2017.

[10] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologie. Stroudsburg, PA, USA: ACL, 2011: 142-150.

[11] KIM Y. Convolutional neural networks for sentence classification[EB/OL].[2019-07-10]. arXiv: 1408.5882v2, 2014.

[12] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

(编辑:刘彦东)