第53卷第5期
2021年10月

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

Vol. 53 No. 5
Oct. 2021

DOI:10.16356/j.1005-2615.2021.05.003

基于相似度的半监督学习工业数据分类算法

孙栓柱[1,2],陈　广[1],高　阳[1],孙　彬[2],李　逗[2],杨晨琛[2]

(1.南京大学计算机科学与技术,南京 210023;2.江苏方天电力技术有限公司,南京 211102)

摘要: 针对现实场景中大量无监督数据无法有效利用的特点,提出了一种基于数据相似度匹配的半监督学习算法。该方法结合一定的先验知识,通过无监督学习的方式,计算未标记数据与少量有标记数据之间相似度,从而对少数类样本进行扩充。利用构造后的数据集进行模型训练,从而提高模型对于少数类的识别效果。该方法能有效改进分类任务中数据分布不平衡及标记困难的问题,在一组基于真实场景下的电力传感器检测数据分类任务中取得了较好的少数类识别效果。通过对比传统以及半监督的多种分类算法,该方法虽然在准确率上低于传统方法,但是在召回率与F_1值的表现上超越传统方法。

关键词: 数据分类;半监督学习;相似度;不平衡学习;不平衡数据分类

中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1005-2615(2021)05-0677-07

Semi-supervised Learning Industrial Data Classification Algorithm Based on Similarity

SUN Shuanzhu[1,2], *CHEN Guang*[1], *GAO Yang*[1], *SUN Bin*[2], *LI Dou*[2], *YANG Chenchen*[2]

(1. Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China;
2. Jiangsu Frontier Electric Technology Co. Ltd., Nanjing 211102, China)

Abstract: Despite their prior knowledge, a large amount of unsupervised industrial data cannot be effectively exploited in real-world applications regions. In this paper, we propose a semi-supervised learning algorithm based on similarity measurement. This method combines specific prior knowledge and unsupervised learning to calculate the similarity between unlabeled data and a small amount of labeled data to augment the minority samples. By improving the classification effect of the model on the minority class, we can mitigate sample imbalance in training phrases and marking difficulty. Empirically, a good minority recognition effect has been achieved in a series of power-sensor detection classification tasks. Compared with state-of-the-art methods, including the traditional and semi-supervised methods, the recall rate and the F_1 value comprehensively exceed the traditional ones.

Key words: data classification; semi-supervised learning; similarity; unbalanced learning; unbalanced data classification

数据分类问题是数据挖掘领域的典型问题,一个表现良好的分类模型,往往离不开充分的有监督数据的支持。然而在现实的应用场景之中,受限于数据标记的难度以及正负样本分布比例等一系列问题,含标记的有监督数据往往十分有限,并且这有限的标记据还会存在类别标签分布不平衡的情况。所以对于此类数据,基于其数据特点,如果将传统的分类算法应用于此类任务之中,往往会过拟

基金项目: 江苏方天电力技术有限公司科技基金(KJ201919)资助项目。

收稿日期: 2020-09-21;**修订日期:** 2021-01-05

通信作者: 孙栓柱,男,教授级高级工程师,E-mail:15905166613@139.com。

引用格式: 孙栓柱,陈广,高阳,等. 基于相似度的半监督学习工业数据分类算法[J]. 南京航空航天大学学报,2021,53(5):677-683. SUN Shuanzhu, CHEN Guang, GAO Yang, et al. Semi-supervised learning industrial data classification algorithm based on similarity[J]. Journal of Nanjing University of Aeronautics & Astronautics,2021, 53(5):677-683.

合于标记数据中的多数类，难以识别出少数类，从而无法取得让人满意的效果。

针对一组给定的数据 Data$=\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$，从而预测一个离散值的任务被称为“分类”，依照输出分类目标类别的不同，分类任务可以被分为“二分类”与“多分类”。数据分类任务的目标便是在于建立输入空间X到输出空间Y之间的映射$f: X \mapsto Y$。

数据分类任务往往需要有监督数据的支撑，有监督数据质量的高低很大程度上影响着模型分类的效果。对于大多数常见的公开数据集，一般拥有着大量的数据标记样本，且样本在类别上的相对分布比较均匀。然而在现实的某一应用场景之中，数据的质量与数量往往是机器学习任务所要面临的第一个难题，这很大程度上制约了模型的效果。

现实场景下分类任务的的数据一般包含以下难题：

(1) 数据的有监督信息有限。现实场景中的机器学习任务，很可能积攒了十分丰富的历史数据，然而这些数据中包含标记的数据十分有限，所以从有监督学习的角度来看，大量数据无法构造监督信息，从而造成模型仅能从有限的数据中学习特征。

(2) 数据的类别分布不平衡。在某一领域的数据之中，数据在类别上的分布可能存在着不平衡的问题。在这样的数据集中，不同标签下的数据量之间不成正比，与此同时在类别间数据量的比例上，以一个二分类任务而言，负正样本之间的比例可能高达999:1，这样的数据往往难以实现对占比较少的数据类别进行识别。

(3) 数据的标记内容具有强领域性。现实中的数据标注任务很可能需要领域性很强的专业知识，猫狗图片数据的标注对于绝大多数人而言都可胜任，但是利用X光片判断病患是否患有癌症，却只有受训多年的肿瘤内科医生才能胜任。此类数据标注的强领域性，制约了该类任务只能在小样本的数据上展开工作，从而限制了模型分类的效果。

本文主要讨论一种基于相似度的半监督分类算法，主要针对有监督标记数据有限，标记数据类别不均衡以及标记内容领域性强的场景。通过计算无标记数据与有标记数据相似度的方式扩充少数类集合，利用半监督学习的方式提高模型对于少数类的分类识别效果。

1 相关研究现状

1.1 不平衡学习

对于部分的数据而言，数据分布在数据的类别上往往不是均衡的，对于那些类别严重失衡的问题常被定义为不平衡学习[1]问题。不平衡学习是指数据集合在类别分布上的不平衡。以分类任务为例，数据中某一类别的数据占总数据中的比例远远高于其他类别[2]，对于这样的数据，占比较高的类别被称为多数类，占比较低的被称为少数类。不平衡数据分类任务广泛地存在与生产与生活中，这种比例失衡的程度很可能达到1 000:1，甚至10 000 000:1。例如，某些罕见疾病的病例数量远远小于其他疾病，电厂环保数据监测传感器异常点的数量远远小于正常点的数量，地震油气勘探领域有油气的地震数据远远小于无油气地震数据的数量。

针对不平衡学习分类任务的特点，主要从以下两个方进行解决：(1)通过调整数据分布的方法进行优化；(2)通过改进模型算法的方式进行优化。

通过改变数据分布的优化方法，主要是通过数据采样的方式，利用一定的手段对数据类别比例进行调整，这样将在一定程度上缓解数据不平衡的问题，使得数据的分布趋向于平衡状态，数据采样一般分为2种方法：(1)对不平衡数据集中的少数类$S_{\min}$进行重采样；(2)对多数类$S_{\max}$中的样本欠采样[3]。前者主要目的在于增加$S_{\min}$的样本，一般采用复制$S_{\min}$的方式，但是这在一定程度上造成了$S_{\min}$的样本冗余。后者一般采用移除某些$S_{\max}$数据的方式，其主要目的在于降低$S_{\max}$的比例，但是这种方式很有可能会在移除数据的过程中造成某些数据信息的丢失。

在数据分布调整上，有Chawla等提出的一种通过创造合成$S_{\min}$样本来实现对少数类过采样的方法，称之为SMOTE(Synthetic minority over-sampling technique)[4]方法，其主要思想是于每一个样本$x_i \in S_{\min}$计算x_i与$S_{\min}$中其他样本之间的欧氏距离，并返回x_i的k个最近值。随后根据全体样本集合S的样本不平衡情况，从少数类集合中挑取2个相邻的样本x及$\hat{x}$，并利用$x_{\text{new}} = x + \text{rand}(0,1)(\hat{x} - x)$的计算方式构造新数据。在SMOTE方法的基础上，Chawla等将Boosting方法结合起来，提出了一种SMOTEBoost[5]方法，通过将SMOTE方法应用于每一个Boosting过程中，对少数类$S_{\min}$中构造新的样本，间接改变了样本分布的不均衡。SMOTE算法从本质上来看是一种过采样的方法，它克服了过采样的一些缺点，通过

数据增强的方法增加了原始数据。除此之外，改进的算法还包括 Borderline-SMOTE 算法[6]与 ADASYN 算法[7]。

在算法模型上，Domingos 等提出了一种基于代价敏感的学习算法[8]，对于一个不平衡数据集，其不同的类别 i 与 j，Cost(i,j)表示类别 i 划分为类别 j 模型所返回的损失。针对少数类别 Min，与多数类别 Max。一般情况下 Cost(Max, Min) > Cost(Min, Max)，因为少数类的样本数量较少，少数类误分类所导致的代价往往要高于多数类的误分类。代价敏感型学习的关键是应用代价敏感矩阵[9-10]，其核心思想是针对数据分布的特点以及一些先验知识，对于不同的分类结果，返回不同的损失，加强模型对于少数类的学习效果。

1.2　半监督学习

半监督学习[11-13]的核心思想在于充分利用有限的有标记数据，结合大量的无标记数据进行模型训练，从而缓解有标记数据样本不充分导致的模型效果表现较差的问题。自 20 世纪 90 年代起，在自然语言处理与计算机视觉需求的驱使下，半监督学习取得了长足的发展，半监督学习的思想发端于 Merz 等[14]。半监督分类学习中，Blum 和 Mitchell 从基于差异的视角，提出协同训练方法[15]，针对有标记的数据从不同的视图，构造不同的属性集，随后利用这些集合进行训练，从而得出不同的模型。然后利用上述模型对大量的无监督数据进行预测，并将置信度较高的结果交叉输入到其他模型之中，反复迭代训练，直到满足条件。该方法表明当训练数据的视图充分冗余时，无标记数据在不同学习器上的一致性能达到最大化，可以有效地降低误分类。从判别式方法的角度，半监督学习利用最大间隔算法[16]训练模型，从而学习得出无标记数据与有标记数据之间的划分边界。基于图的半监督分类方法主要通过基于流形假设[17]原理，构建数据集中样例之间的图关系，随后基于图之间的关系实现标记数据的有监督信息向无监督数据的传播。首先基于图的方法会选择合适的距离计算样例之间的距离，如欧氏距离、切比雪夫距离和马氏距离等。随后根据前述计算所得的距离选择合适的连接方式，构造样例之间的连接图。在图构造完成的基础上利用核函数计算连接边的权值，并利用这个权值衡量两个连接点之间的相似度。

2　半监督相似度量工业数据分类算法

2.1　问题分析

对于一个分类任务而言，以二分类任务为例，一个分类效果良好的分类器往往需要充分利用向好的正负样本进行学习，从而学习出正负类别中的特征 θ。但是基于前文所述，在现实的应用场景之中，经常存在数据标注难度大、数据样本分布不均衡以及标记信息有限等诸多问题，以上问题所导致的直接影响便是用于学习的标记样本其分布上存在着不均衡。

基于样本类别分布不均衡的数据所训练得到的分类器，往往会过拟合于不均衡数据集中的多数类 S_{max}，从而难以识别少数类。这种情况下仅从准确率的视角衡量模型的效果便不够客观，因为数据集中多数类样本充分，可供学习的数据众多，分类器便能够充分学习出多数类中的特征 θ_{max}。但是这样的分类器在本质上过拟合于多数类 S_{max}，分类器几乎无法识别出所有的少数类 S_{min}。在现实的应用场景之中，对于不平衡数据而言，相比识别常见的多数类，识别出不平衡数据中的少数类 S_{min} 往往更具有价值。

本文所要处理的分类任务来自于某一工业领域，在某一区域范围之内均匀散布了几十万个传感器，其中绝大部分数据是无标记数据。有标记的数据划分为两类结果，无显示数据 N 与有显示数据 P，以及半监督数据 P'，其中有显示数据 P 为主要的识别目标，D 为半监督数据的筛选范围，如图 1 所示。基于其业务特点，其所有的标记数据中，有显示数据 P 远远小于无显示数据 N，而且数据的标签信息的获取，需要专业的工作人员现场在每一个传感器安置点进行施工采样，验证传感器放置点的现场状况，才可以判断该监测点的标记为有显示点 N，还是无显示点 P，所以数据的标记信息十分有限，仅仅几百条。该工业领域的业务人员表示，有显示数据与无显示数据往往是以范围的形式存在，

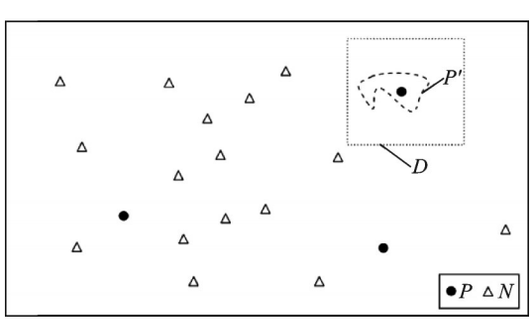

图 1　数据采样分布图

Fig.1　Data sampling distribution diagram

但是在有、无显示的范围内仅仅以几个传感器的特征尤为明显，并基于这些少数的传感器进行标记正负情况。

此工业数据分类任务存在以下难题：(1)样本数据众多，但含有标记信息样本较少；(2)标记数据中正负样本比例差距大；(3)数据标记难度大。

本文提出一种基于相似度度量的半监督分类算法。其主要的方式是，围绕任务目标数据集中的少数类 $S_{\min}$ 中的样本，针对其中的每一个少数类正例样本 $s_{\min}$，在 $s_{\min}$ 周围限定的一个区域，该区域范围内包的无标记数据集合为 $D=\{s_1^D, s_2^D, \cdots, s_n^D\}$，对与区域内的某一个样本 s_i^D，其并不存在标签，随后对 D 内的所有数据进聚类。其核心思想是针对任务目标数据中样本数量比例差距较大的特点，缓解分类器在训练过程中过拟合于占比较大类别的数据所导致的问题。

针对上述任务描述，在同业务人员的交流中得知，标记为正例的数据其周围的数据大概率也为正类，相同类别的数据之间的相似性较高，随后从聚类结果中挑选与区域 D 内正例标记数据 P 最为相似的类别集合 P'，将 P' 其作为可信正例集合，并以此扩充正例样本，缓解数据标记集合中样本分布不均衡的情况，最后利用扩充集合中的数据进行模型训练，并得出分类器。

2.2 基于K-means聚类相似度扩展正例集合

基于前文所述，有显示数据 P 为数据集合中的少数类，由于标记数据的难度较大，所以训练集中少数的有显示数据 $P=\{p_1, p_2, \cdots, p_n\}$，无法充分反映少数类数据特征在全局状态下对于全体少数类集合 P 的分布。因此很有必要针对 P 进行扩充。

本文使用K-means[18]聚类算法对可信正类数据 P 周围的无标记数据集合 D 进行聚类，依赖半监督学习中的平滑假设[19]与聚类假设[20]。所谓的平滑假设即位于数据稠密的区域中，距离相近的样例，大概率拥有相同的类标签。所谓的聚类假设，即处于相同类簇样例下的样例，具有相同类标签。

存在两个问题需要明确：

(1) 如何制定一套机制，以确定K-means算法中 k 的取值。对于一个有监督分类任务而言，数据需要被划分的类别是明确的。如图2所示，对于需要施加K-means算法的数据集合 $D=\{s_i^D\}_{i=1}^n$，其标签集合也为 $\text{Label}=\{P, N\}$，类别为2。如果直接设置 k 的取值为2，那么数据集合 D 将会很粗略地被划分为两个类别，基于聚类假设的原理，会将一个数量较大的类簇划归为少数类，这种粗糙的少数类数据扩充方式，无疑会增加少数类数据中的噪声，以此数据进行模型训练将会得到一个表现较差的分类器。

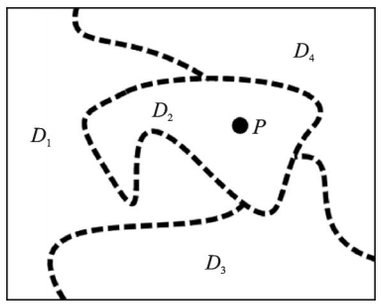

图2 围绕可信数据 P 的聚类结果示例图

Fig.2 Example graph of clustering results around trusted data P

(2) 如何建立一个方法，以度量K-means算法聚类所得的 k 个类别中与可信少数类数据 P 之间的相似度[21]，并从K-means聚类结果 $D=\{D_i\}_{i=1}^{i=k}$ 中，挑选出与 D 范围内可信数据 P 之间相似度最近小的类别。常见的方式是以距离方式进行度量，本文选用最经典欧式距离作为数据的度量方式来计算聚类样本与标记样本之间的相似度。对于一个 k 值下的K-means聚类结果 $D_i=\{d_j^i\}_{j=1}^{j=n}$，其相似度计算方式为

$$\text{Distance}(P, D_i)=\sqrt{\sum_{j=1}^{n}(d_j^i-P)^2} \tag{1}$$

针对以上问题，本文提出了一种基于K-means聚类的相似度收敛算法。通过设置一组逐渐递增的 k 值，随着 k 值的递增加，K-means聚类得到的类别逐步精细，当 k 个类别中与标记数据 P 的距离开始收敛的时候，停止 k 值的递增，并将该类别作为可信正例集合，其整体流程如图3所示。

以图4为例，其为某一标记节点周围，距离其最近的K-means聚类数据分布变化图，其中，k 表示K-means聚类算法中 k 的取值，num表示K-means聚类结果中距离标记节点最近集合中数据的数量，d 表示该集合中距离标记节点的平均距离。可以

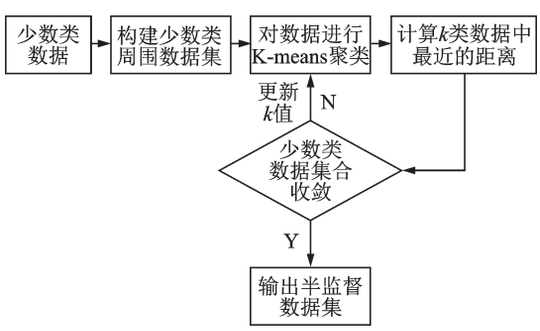

图3 半监督数据扩充流程图

Fig.3 Flow chart of semi-supervised data expansion method

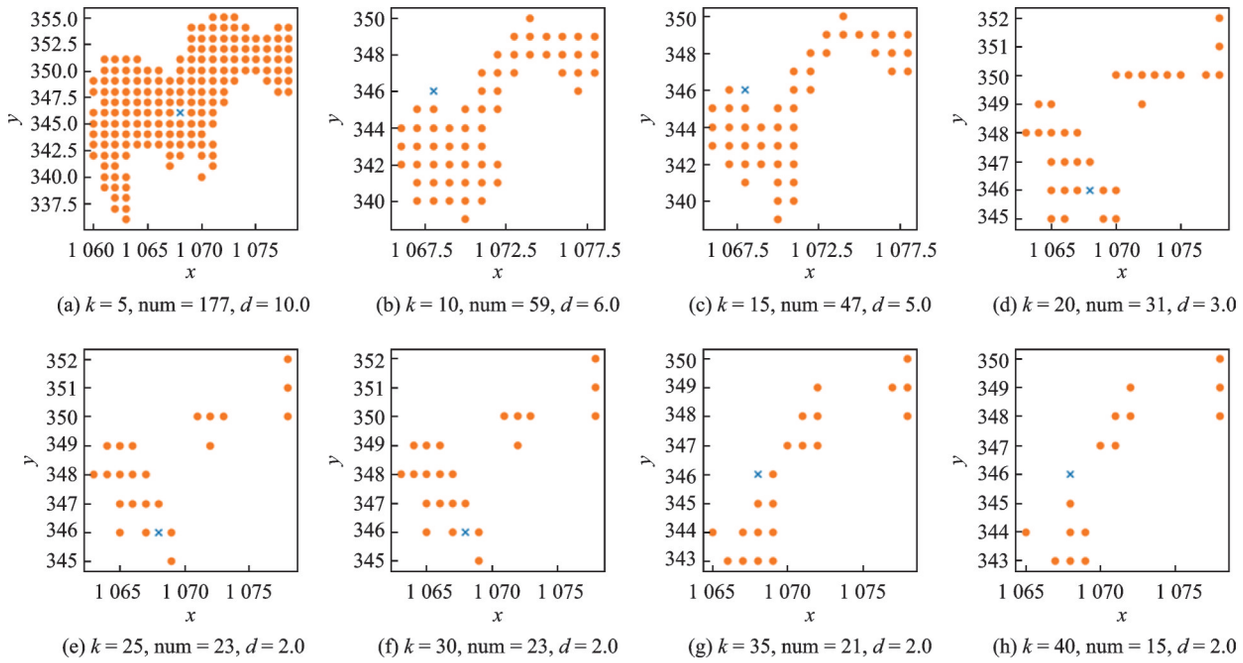

图 4　可信数据 P 点 K-means 聚类变化图

Fig.4　K-means cluster change graph of trusted data point P

看到距离最近的数据始终围绕在可信节点的周围。图 5 展示了随着 K-means 中 k 值变化过程中，距离标记节点最近的聚类的可信样本数量变化情况，以及与标记节点的距离变化情况，可以看到距离与数量都是逐步下降并最终收敛。

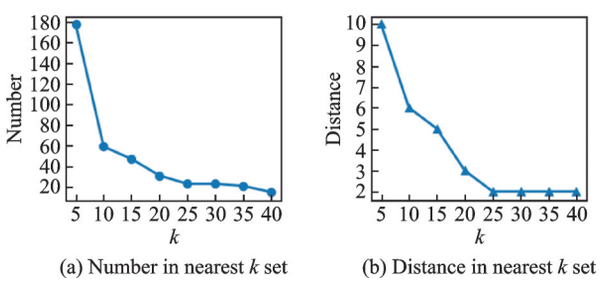

图 5　聚类数目以及距离变化图

Fig.5　Cluster number and distance change graph

可以看到，利用上述方法可以得到稳定的且距离不平衡数据中少数类 $S_{\min}$ 最近的数据集合，这些原本没有标记的数据将作为半监督学习中的训练数据。

算法 1　基于 K-means 聚类的相似度收敛算法

输入：$D, P, K_{\text{values}} = \{k_i\}_{i=1}^{i=n}$

输出：　$D_i|k_i$

```
Initialize i = 1, minDis = +∞, D^k = ø, λ
for each k_i ∈ K_values do
  D_k = K-means(D, k_i)
  for each D_k^i ∈ D_k do
    currenDis = Distance(P, D_k^i)
  if |currentDis − minDis| ≤ λ then
    return k, i, D_k^i
  else
   if currentDis ≤ minDis then
     currentDis = minDis
   else
     minDis = currentDis
   end if
  end if
 End for
End for
```

2.3　利用多种分类算法进行数据分类

为了有效验证利用 K-means 聚类，并以此取得最相似数据，从而进行验证并比较半监督分类学习的算法效果。本文利用多种分类算法进行验证，包括一系列浅层模型与深度模型。其中浅层模型包含以下算法：KNN（K-nearest neighbor）决策树、SVM（Support vector machine）和 LR（Logistic regression）分类器。深度学习分类器包括：全连接神经网络、循环神经网络与长短时神经网络。

3　实验及结果分析

3.1　实验评价

为了有效地评价模型的分类效果，本文主要应用到了准确率（Accuracy）、召回率（Recall）以及标准的 F_1 度量，式（2～4）中其他变量含义为：TP 表示数据自身为正例并被识别为正例；FN 表示数据自身为正例但是被识别为负例；FP 表示数据自身为负例但被识别为正例；TN 表示数据自身为负例并被识别为负例；Precision 表示精确率。其中准确

率主要衡量分类器的预测结果中有多少是分类正确的。但是基于不平衡数据集中样本不均衡的问题,仅仅使用准确率无法客观的评价模型的效果,所以模型还会参考召回率这一指标,以衡量对于少数类的分类效果。最后利用F_1值综合评价分类器的效果。

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{3}$$

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

3.2 实验数据集

本文所应用到的数据集来自于某一工业领域,共包含传感器采集的数据221 121条,每一条数据包含30个特征,其中包含标记的数据仅有480条,标记数据中正例数据31条,其余皆为负例。在此标记数据的基础上,利用前文所述的方法,围绕着正例标记数据构造可信正例数据157条。以上为本实验所应用到的数据集。

3.3 实验结果及分析

在实验过程中,采用如下方式进行实验,其中对于标记数据集,将其中的80% 作为训练集,20%作为测试集。对于可信数据集,则只将其添加到训练集中,构造半监督训练集。实验过程中将分别利用训练集与半监督训练集进行模型训练,得出普通的分类器与半监督分类器。然后利用测试集评价上述两组分类器的效果。为准确衡量本文算法的效果,利用了多个分类器来评判算法的效果。

从表1可以看出,传统方法的准确率明显高于半监督方法,但是传统方法中除决策树模型之外,其他模型的召回率与F_1值均为0,这表明基于传统方法训练得出的分类器基本上过拟合于不平衡数据集中的多数类,并且模型不具有少数类分类的能力,而不平衡数据集中的少数类,往往是该类型分类任务重关注的重点。对比于半监督方法,模型的召回率与F_1值均有所提高,这表明半监督方法相对于传统模型,在对于数据集中的少数类而言其识别率有所提高。

表1 传统与本文半监督方法对比

Table 1 Comparison of traditional and semi-supervised method

分类	方法	准确率	召回率	F_1
传统方法	KNN	0.83	0.00	0.00
	决策树	0.76	0.28	0.28
	SVM	0.83	0.00	0.00
	LR	0.83	0.00	0.00
	全连接	0.83	0.00	0.00
	RNN	0.83	0.00	0.00
	LSTM	0.83	0.00	0.00
半监督方法	KNN	0.62	0.42	0.27
	决策树	0.74	0.57	0.59
	SVM	0.37	0.42	0.18
	LR	0.37	0.42	0.18
	全连接	0.44	0.42	0.20
	RNN	0.67	0.57	0.36
	LSTM	0.60	0.42	0.26

4 结 论

本文提出了一种基于半监督学习的工业数据分类算法,该方法针对需某类传感器数据分类任务,在其标记信息少、标记难度大以及正负样本分类不均衡的情况下,创新地提出利用K-means聚类算法,将围绕少数类标记数据中的数据进行无监督聚类,基于聚类数据与少数类标记数据的相似度度量,构造一批可信正例半监督数据,并将其用于模型训练,从而得到半监督分类器,经多种分类模型进行半监督训练测试验证,虽然模型的准确率受到了一定的影响,但是模型在召回率与F_1值的表现上明显优于传统的方法,这表明本文方法有效地识别出了不平衡工业传感器数据数据集中重点需要识别出的少数类,在真实的生产与应用环境中具有一定的应用价值。

参考文献:

[1] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[2] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6.

[3] DRUMMOND C, HOLTE R C, et al. C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling[C]//Proceedings of Workshop on Learning From Imbalanced Datasets II: volume 11. [S.l.]: Citeseer, 2003: 1-8.

[4] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[5] CHAWLA N V, LAZAREVIC A, HALL L O, et al. Smoteboost: Improving prediction of the minority class in boosting[C]//Proceedings of European Conference on Principles of Data Mining and Knowledge

Discovery. [S.l.]: Springer, 2003: 107-119.

[6] HAN H, WANG W Y, MAO B H. Borderline-smote: A new over-sampling method in imbalanced data sets learning[C]//Proceedings of International Conference on intelligent Computing. [S.l.]: Springer, 2005: 878-887.

[7] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of 2008 IEEE International Joint Conference on neural Networks (IEEE World Congress on Computational Intelligence). [S.l.]: IEEE, 2008: 1322-1328.

[8] DOMINGOS P. Metacost: A general method for making classifiers cost-sensitive[C]//Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Francisco, CA, USA: ACM, 2001.

[9] ELKAN C. The foundations of cost-sensitive learning[C]//Proceedings of International Joint Conference on Artificial Intelligence. [S.l.]: Lawrence Erlbaum Associates Ltd, 2001: 973-978.

[10] ZADROZNY B, LANGFORD J, ABE N. Cost-sensitive learning by cost-proportionate example weighting[C]//Proceedings of Third IEEE International Conference on Data Mining. [S.l.]: IEEE, 2003: 435-442.

[11] ZHU X J. Semi-supervised learning literature survey[R]. [S.l.]: University of Wisconsin-Madison Department of Computer Sciences, 2005.

[12] ZHU X, GOLDBERG A B. Introduction to semi-supervised learning [J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3 (1): 1-130.

[13] 刘建伟, 刘媛, 罗雄麟, 等. 半监督学习方法[J]. 计算机学报, 2015,38(8): 1592-1617.
LIU Jianwei, LIU Yuan, LUO Xionglin, et al. Semi-supervised learning method[J]. Journal of Computer, 2015,38(8): 1592-1617

[14] MERZ C J, CLAIR D S, BOND W E. Semi-supervised adaptive res- onance theory (smart2)[C]//Proceedings 1992 IJCNN International Joint Conference on Neural Networks. [S.l.]: IEEE, 1992: 851-856.

[15] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. Madison, Wisconsion: ACM, 1998.

[16] HSIAO J Y, TANG C Y, CHANG R S. An efficient algorithm for finding a maximum weight 2-independent set on interval graphs[J]. Information Processing Letters, 1992, 43(5): 229-235.

[17] LIN T, ZHA H. Riemannian manifold learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(5): 796-809.

[18] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained K-means clustering with background knowledge[C]//Proceedings of ICML. [S.l.]: ACM, 2001: 577-584.

[19] LU S, PEREVERZEV S V, RAMLAU R. An analysis of tikhonov regularization for nonlinear ill-posed problems under a general smoothness assumption[J]. Inverse Problems, 2006, 23(1): 217.

[20] RIGOLLET P. Generalization error bounds in semi-supervised classification under the cluster assumption[J]. Journal of Machine Learning Research, 2007, 8: 1369-1392.

[21] GUO G D, JAIN A K, MA W Y, et al. Learning similarity measure for natural image retrieval with relevance feedback[J]. IEEE Transactions on Neural Networks, 2002, 13(4): 811-820.

(编辑:刘彦东)