计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200345
58(9): 2052-2061, 2021

基于宏微观因素的概率级别的车辆事故预测

张力天[1,2] 孔嘉漪[1,2] 樊一航[1,2] 范灵俊[2,3] 包尔固德[1]
[1](北京交通大学软件学院 北京 100044)
[2](中国科学院计算技术研究所信息技术战略研究中心 北京 100190)
[3](贵阳市大数据产业集团有限公司 贵阳 550081)
(remilia@bjtu.edu.cn)

Car Accident Prediction Based on Macro and Micro Factors in Probability Level

Zhang Litian[1,2], Kong Jiayi[1,2], Fan Yihang[1,2], Fan Lingjun[2,3], and Bao Ergude[1]
[1](*School of Software Engineering, Beijing Jiaotong University, Beijing* 100044)
[2](*Information Technology Strategy Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[3](*Guiyang Big Data Industrial Group Co. Ltd., Guiyang* 550081)

Abstract Car accident prediction is an important problem to study for avoiding the accidents. Previous studies make the prediction for a car based either on macro factors such as geography, environment and traffic or on micro factors such as car and driver behaviors. There is rarely a study combining the two types of factors because it is difficult to collect the two types of data at the same time. However, car accidents usually result from both of the two types of factors. In addition, the current researches predict whether an accident will happen or not. There is rarely a study providing a more accurate accident probability because there is no probability label for use in the collected data. However, such a probability is useful to notify the driver in different warning levels. The OSU(Ohio State University) accident dataset of macro factors published in 2019 has some identical characteristics with the FARS (fatality analysis reporting system) dataset of macro factors and SHRP2(strategic highway research program 2) dataset of micro factors, and thus provides an opportunity to fuse them. Therefore in this paper, we obtain a dataset of both macro and micro factors. In the dataset, accident data (positive data) is fused from the OSU and FARS datasets, as well as Sim-SHRP2(simulated strategic highway research program 2) similar to the SHRP2 dataset, while safe-driving data (negative data) is obtained by ourselves driving a car. In addition, since the obtained dataset does not have any probability label, we also design a probability-level unsupervised deep learning framework to predict the accurate probability. The framework iteratively generates accurate probabilities from the obtained dataset, and is trained with the generated probabilities. The experimental results indicate our framework can predict car accidents with the obtained dataset sensitively and accurately.

收稿日期:2020-05-25;**修回日期**:2021-02-04
基金项目:中央高校基本科研业务费专项资金(2019JBM073);北京市自然科学基金项目(4192044);国家自然科学基金重大培育项目(91646127);中国工程院重大咨询课题(2018-ZD-02)
This work was supported by the Fundamental Research Funds for the Central Universities (2019JBM073), the Beijing Natural Science Foundation (4192044), the Major Cultivation Project of the National Natural Science Foundation of China (91646127) and the Major Consulting Project of Chinese Academy of Engineering (2018-ZD-02).
通信作者:包尔固德(baoe@bjtu.edu.cn)

Key words car accident; accident prediction; macro and micro factors; deep-learning framework; deep-SVDD algorithm

摘 要 车辆事故预测是避免道路车辆事故发生的重要研究课题.以往的研究使用的事故数据集只包含地理情况、环境情况、交通情况等宏观因素,或者只包含车辆行为和驾驶员行为等微观因素.因为很难收集到同时包含2类因素的事故数据集,很少有研究将这2类因素结合起来,然而车辆事故往往是两者共同作用的结果.此外,在收集到的数据中没有可以用于预测的事故发生概率标签,所以目前多数的研究关注点只是在于事故是否发生而不能得到准确的概率值.然而在实际应用场景下,驾驶员需要的是不同级别的危险预警信号,而这种信号正是应该由事故概率值决定的.2019年发布的事故宏观因素数据集OSU(Ohio State University)与宏观因素数据集FARS(fatality analysis reporting system)和微观因素数据集SHRP2(strategic highway research program 2)都具有一些相同的特征,为它们的融合提供了机遇.因此,首先得到了一个同时包含宏观和微观因素的数据集,其中事故数据(正样本)融合自OSU、FARS数据集,以及与SHRP2分布相同的数据集Sim-SHRP2(simulated strategic highway research program 2),而安全驾驶数据(负样本)则由自己驾驶汽车获得.然后,针对收集到的数据中没有概率标签的问题,还设计了一个概率级别的无监督深度学习框架来预测准确的概率值,该框架使用迭代的方式为数据集生成准确的概率标签,并使用这些概率标签来进行训练.实验结果表明,该框架可以使用所得到的数据集来灵敏而准确地预测车辆事故.

关键词 车辆事故;事故预测;宏微观因素;深度学习框架;deep-SVDD算法

中图法分类号 TP391

现阶段,驾车出行已经变成人们日常生活中的一部分.车辆给人们的生活带来了很多便利,但也带来了很多交通安全隐患.根据美国国家公路交通安全管理局(National Highway Traffic Safety Administration, NHTSA)发布的《机动车碰撞概述》,美国每分钟大约发生10起车辆事故,每3分钟就有1人因车辆事故而丧生[1].根据世界卫生组织2018年发表的《全球道路安全状况报告》,全球有超过135万人死于道路车辆事故.车辆事故问题已经变成世界关注的焦点问题.预防车辆事故的方法有很多,政策方面包括提高驾驶员的安全意识、加大处罚力度和更新交通系统设备,而技术方面主要是在事故发生前预测事故并通知警示驾驶员.

在事故预测中,获取车辆事故数据是很重要的环节.在以往的研究中,这些数据集包含宏观因素[2-4]或微观因素[5-8].宏观因素包括地理情况(如公路和街道)、环境情况(如湿度和能见度)和交通情况(如交通速度和密度).这些包含宏观因素的数据集通常是从当地交通部门收集到的,例如RCP(part of the Ministry of Justice and Public Order of the Republic of Cyprus),CIAIR(Center for Integrated Acoustic Information Research), SPD (Seattle Police Department)数据集[9-11].微观因素包括事故发生前的车辆行为(如速度和加速度)和驾驶员行为(如眼动和手势).这些数据集通常使用驾驶模拟器收集,例如STSIM,SCANeR,VANET数据集[7,9].在以往的研究中,大部分数据集只包含宏观或微观其中一种因素.究其原因,是宏观因素和微观因素难以同时收集.然而,车辆事故往往是两者共同作用的结果[1].虽然Gite和Agrawal使用了2类因素的数据进行事故预测,但数据集包含的特征数量非常少[5].

事故预测的另一重要环节是用算法处理数据,而当前有很多种预测车辆事故的方法.在深度学习出现之前,人们通常使用神经网络[10,12-14]或统计方法,如条件logistic模型等[15-18].目前,很多研究者使用LSTM处理时间序列事故数据[5],或使用CNN和DNN处理非时间序列数据[2,4,19],来预测驾驶员的危险行为或者预测出事故发生的时间和地点.因为收集到的数据中没有可以用于预测的事故发生概率标签,以往的研究通常只是计算出一个0/1二分类预测结果,即仅仅判断事故能否发生,但这样的结果不够实用,驾驶员更希望得到的是不同级别的危险预警信号[20].如果可以预测出事故概率值,就能更好地满足驾驶员的需求.

2019年发布的事故宏观因素数据集OSU(Ohio State University)与宏观因素数据集FARS(fatality

analysis reporting system)和微观因素数据集 SHRP2 (strategic highway research program 2)都具有一些相同的特征,为它们的融合提供了机遇[21-22].因此在本研究中,我们首先得到了一个同时包含宏观和微观因素的数据集.对于正样本数据,我们融合了包含宏观因素的 OSU 和 FARS 数据集,以及包含微观因素且与 SHRP2 分布相同的 Sim-SHRP2(simulated strategic highway research program 2)数据集.而负样本数据,则是在美国加利福尼亚州和华盛顿州自行驾驶汽车收集而来的.此外,融合后的数据集只包含事故是否发生的 0/1 标签而不包含事故发生的概率标签,所以我们还设计了一个深度学习框架来预测车辆事故发生的准确概率值.该框架参考了 Zhang 等人[23]提出的算法,是一个无监督的深度学习框架,但是它可以使用迭代的方式为数据集生成准确的概率标签,并使用这些概率标签来进行训练.由于时间和资金的限制,负样本数据量约为正样本数据量的 1/5,所以在该框架中,我们使用了 deep-SVDD 算法,该算法能够很好地解决数据不均衡的问题[24].

综上所述,本文的贡献主要在实际应用上,解决了 2 个问题:1)车辆事故是宏观因素和微观因素共同作用的结果[1],但是因为宏微观因素难以同时收集,所以以往的大部分车辆事故数据集只包含宏观或微观其中一种因素.针对这一问题,我们融合了 2019 年发布的事故宏观因素数据集 OSU 与宏观因素数据集 FARS 和微观因素数据集 SHRP2,得到了包括宏微观因素的数据集.2)不同层级的事故发生概率的预警可以有效地避免交通事故,并且也符合驾驶员的需求[20],但是因为收集的数据集中没有可以用于预测事故发生概率的标签,所以以往的大部分研究只得到事故是否发生的二分类结果.针对这一问题,我们设计了一个概率级别的无监督深度学习框架,使用迭代的方式为数据集生成准确的概率标签,并使用这些概率标签来进行训练,从而可以预测出准确的事故发生概率值.

1 方法和数据

1.1 总体介绍

为了得到完整而足量的训练数据,我们首先融合了 3 个包含正样本的车辆事故数据集,然后将它们与一个包含负样本的安全驾驶数据集相结合,最终获得了可用于训练的完整数据集.正样本的 3 个数据集分别是 OSU,FARS,Sim-SHRP2,其中前 2 个数据集包含事故的宏观因素,后一个包含事故的微观因素.OSU 数据集是从俄亥俄州立大学收集的涵盖美国 49 个州的全国性车辆事故集合,其中包含车辆事故的地理情况和环境情况相关特征.FARS 数据集是在美国进行的全国性年度事故普查的数据,其中包含交通情况相关特征.Sim-SHRP2 数据集也是美国的全国性事故统计,其中包含车辆事故中的车辆行为和驾驶员行为相关特征.负样本数据集则是我们与腾讯自动驾驶组人员合作,驾驶配备有各类传感器的汽车在加利福尼亚州和华盛顿州收集到的.数据融合过程如图 1 所示:

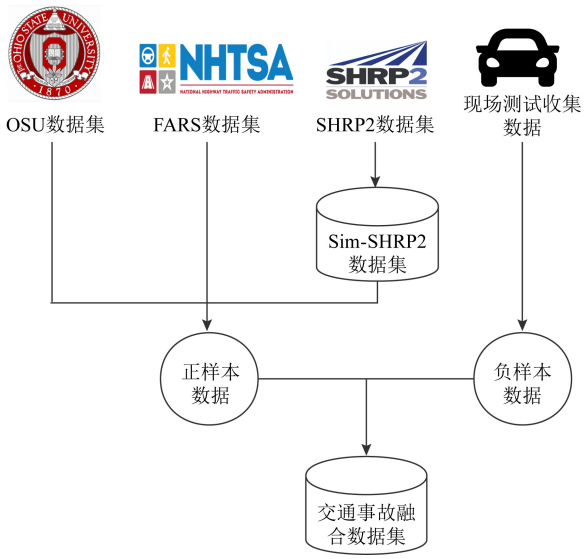


Fig. 1 Illustration of data fusion process

图 1 数据融合流程示意图

在构建好包含正负样本的具有多个特征的数据集后,我们将数据输入深度学习框架中,首先通过无监督算法生成粗糙概率标签,然后使用粗糙的标签迭代地训练深度学习算法,得到更准确的标签,再用所得到的标签进行训练.框架主要分为 3 步:

1) 使用无监督的机器学习算法对数据进行预处理,生成粗略的概率标签;

2) 使用有粗糙概率标签的数据训练有监督的深度学习算法;

3) 使用训练后的深度学习算法进一步处理数据,使生成的概率标签更准确.

步骤 2)和步骤 3)不断迭代进行,直到深度学习算法得到充分训练为止.

因为训练数据只包含事故发生的 0/1 标签,并不包含我们所需要的概率标签,所以我们以迭代的方式逐渐得到准确的概率标签.在本文中,选择的无监

督机器学习算法是 k-means 聚类算法和 autoencoder 自编码器,而有监督的深度学习算法是一种修改的 deep-SVDD 算法.原始的 deep-SVDD 算法为无监督算法,但是可以通过修改损失函数,改为有监督算法.我们选用无监督 deep-SVDD 算法并把它修改为监督学习算法,主要是因为训练数据的正负样本不均衡,而 deep-SVDD 算法可以很好地处理这样的数据.尽管训练数据可以通过多次迭代逐渐获得准确的概率标签,但是在迭代初期的概率标签比较粗略,也比较适合修改为监督算法的无监督学习,而不是一般的监督算法.整体框架如图 2 所示.值得说明的是,尽管该框架为训练数据生成了概率标签,并且使用修改的 deep-SVDD 算法进行了监督学习,但是它使用的训练数据与 k-means,autoencoder 和原始的 deep-SVDD 算法一样,都是没有概率标签的数据,所以是概率级别的无监督学习框架.

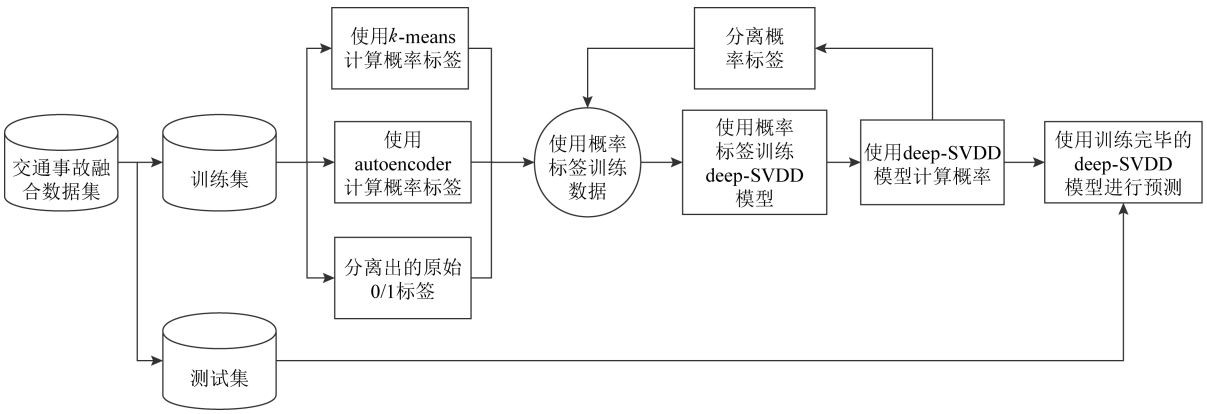


Fig. 2 Illustration of the deep learning framework process

图 2 深度学习框架流程示意图

1.2 事故预测数据收集与融合

1.2.1 正样本数据融合

OSU 数据集由事故发生的地理情况(如"经纬度"和"所在州")、时间、环境情况(如"降雨量"和"能见度")等相关特征构成;FARS 数据集由地理情况、时间、事故严重程度、环境情况、交通情况(如"周边是否有校车"和"周边是否有行人")等相关特征构成;SHRP2 数据集由地理情况、时间、事故严重程度、车辆行为(如"是否超速"和"是否变道")、驾驶员行为(如"是否系安全带"和"是否打瞌睡")等相关特征构成.尽管 FARS 数据集也包含少量驾驶员行为相关特征,但是远不如 SHRP2 数据集全面.我们运行程序来融合 OSU,FARS,Sim-SHRP2 数据集,该程序包括融合数据集和剔除无效数据项等功能.首先,我们根据不同数据集的相同特征来融合数据集.OSU 和 FARS 数据集的融合方式为,如果 2 个事故分别来自于 OSU 和 FARS 数据集,并且它们的经纬度距离之差小于 500 m,时间之差小于 30 min,那么我们就将其视为同一起事故,合并相应的数据项.OSU 和 Sim-SHRP2 数据集的融合方式为,如果 2 个事故分别来自于 OSU 和 Sim-SHRP2 数据集,并且它们的州、其他地理情况(如"是否靠近学校"和"是否靠近交叉路口"等)、时间和事故严重程度完全相同,那么我们就将其视为同一起事故,合并相应的数据项.之后,我们根据相关和相似特征来剔除无效数据项,这些相关和相似特征如 OSU 数据集中的"降雨量"特征和 FARS 数据集中的"地面湿度"特征,以及 FARS 数据集中的"事故前驾驶员行为"特征和 Sim-SHRP2 数据集中的"驾驶员是否分心"特征等.如果某一合并数据项的相关和相似特征不一致,我们就将其视为无效数据并加以剔除.这些不一致一般是因为数据记录时的偏差引起的.其中,对于 Sim-SHRP2 数据集,原始 SHRP2 数据集无法免费下载,但是官方网站提供了所有特征的概率分布,所以我们根据其提供的概率分布模拟了 Sim-SHRP2 数据集.融合后的数据集包含约 850 000 个具有 97 个特征的数据项,这些特征分为 5 个不同类别,分别是:地理情况、环境情况、交通情况、车辆行为和驾驶员行为.

1.2.2 负样本数据收集

为了收集负样本数据,我们使用行车记录仪(Garmin VIRB Ultra 30)、车辆传感器(iVICAR)和视频记录仪(iPhone 11),驾驶汽车(Toyota RAV4)在加利福尼亚州、华盛顿州和周边的多个场景下采集负样本数据.这些场景互不相同,覆盖了特征的多种组合.行车记录仪记录地理情况、环境情况和交通

情况,而视频记录仪记录驾驶员行为,采集到的视频通过手动分析来进一步提取特征.车辆传感器则可以自动记录车辆行为特征,其截图如图 3 所示.由于时间和资金的限制,收集的负样本数据集包含约 170 000 个样本,约占正样本数据集的 1/5.

Fig. 3 Snapshot of the car sensor

图 3 车辆传感器截图

1.3 基于 k-means 和 autoencoder 粗略概率标签生成

我们使用无监督机器学习方法来为训练数据集计算粗略的概率标签.根据 Zhang 等人[23]所述,这里应当使用至少 2 种无监督机器学习方法来保证结果的准确度,所以本项目同时使用了 k-means 算法和 autoencoder.k-means 算法可以将样本划分为 k 个聚类,使得彼此接近的样本分到同一聚类之中.在本项目中,k-means 算法会使训练数据集聚集为 2 个类:聚类 1 主要包含正样本数据,聚类 0 则主要包含剩余的负样本数据.在得到这 2 个聚类后,需要把每个样本到聚类中心的余弦距离归一化到[0,1]之间,然后可以求得样本 i 的粗略概率标签

$$p_i=\begin{cases}1-d_i/2, & i \text{ 属于聚类 } 1,\\ d_i/2, & i \text{ 属于聚类 } 0,\end{cases}$$

其中 d_i 是归一化后的距离.也就是说,对于正样本,距离聚类中心的距离越近,发生车辆事故的概率越大;而对于负样本,距离聚类中心的距离越近,发生车辆事故的概率越小.autoencoder 是一种可以学习无标签训练样本编解码过程的神经网络.本项目中,autoencoder 使用正样本进行训练,这样当输入正样本时,输出的解码数据与原数据差别较小,而输入为负样本时,输出的解码数据与原数据差别较大,就可以对正负样本加以区分[25].每个样本与其解码样本余弦距离的归一化值取负加 1,即是它粗略的概率标签.无论是 k-means 算法还是 autoencoder 算法,都难以直接求得准确的概率标签,所以如 Zhang 等人[23]所述,我们对 k-means 算法和 autoencoder 算法的结果和原始 0/1 标签求平均值,这样所得的结果更加准确,可以取得更好的训练效果.

1.4 用于监督学习的修改的 deep-SVDD 算法

原始的 deep-SVDD 算法是非监督学习算法,在训练阶段不需要数据标签.它用深层神经网络将数据 $\boldsymbol{x}_i$ 映射到 $\phi(\boldsymbol{x}_i;\boldsymbol{W})$,其中正样本映射到中心为 $\boldsymbol{c}$ 且半径为 R 的高维超球面之中,负样本则映射到超球面之外.原始的 deep-SVDD 算法只用正样本进行训练,其损失函数为 $\min\limits_{R,\boldsymbol{W}}R^2+\dfrac{1}{vn}\sum\limits_{i=1}^{n}\max\{0,\|\phi(\boldsymbol{x}_i;\boldsymbol{W})-\boldsymbol{c}\|^2-R^2\}+\dfrac{\lambda}{2}\sum\limits_{l=1}^{L}\|\boldsymbol{W}^l\|^2$.该损数函数试图最小化 3 个目标:超球面的半径 R、映射到超球面外样本到中心 $\boldsymbol{c}$ 的平均距离,以及深度神经网络的权重 $\boldsymbol{W}$.

在本项目中,我们想要在 deep-SVDD 的训练中考虑概率标签,使得它能够在预测中计算出准确的概率值.为此,我们用正负样本同时训练,并在其损失函数中增加了标签相关项 $\dfrac{1}{un}\sum\limits_{i=1}^{n}(2R(1-p_i)-\|\phi(\boldsymbol{x}_i;\boldsymbol{W})-\boldsymbol{c}\|)^2$.修改后的损失函数为 $\min\limits_{R,\boldsymbol{W}}R^2+\dfrac{1}{vn}\sum\limits_{i=1}^{n}\max\{0,\|\phi(\boldsymbol{x}_i;\boldsymbol{W})-\boldsymbol{c}\|^2-R^2\}+\dfrac{\lambda}{2}\sum\limits_{l=1}^{L}\|\boldsymbol{W}\|^2+\dfrac{1}{un}\sum\limits_{i=1}^{n}(2R(1-p_i)-\|\phi(\boldsymbol{x}_i;\boldsymbol{W})-\boldsymbol{c}\|)^2$.在增加的标签相关项中,随概率标签值的减小,样本在分类超球面内的映射位置逐步远离球心,所以概率标签为 p_i 样本的理想映射位置与半径为 R 的超球面中心距离是 $2R(1-p_i)$,而该样本的实际映射位置为 $\phi(\boldsymbol{x}_i;\boldsymbol{W})$,与超球面中心 $\boldsymbol{c}$ 的距离是 $\|\phi(\boldsymbol{x}_i;\boldsymbol{W})-\boldsymbol{c}\|$.该项最小化这 2 个距离,使得 deep-SVDD 能够在预测中计算出准确的概率值,其中 u 是一个调整该项权重的超参数.

1.5 修改 deep-SVDD 算法的迭代训练

1.4 节介绍了对原始 deep-SVDD 算法的修改,而本节介绍对修改 deep-SVDD 算法的迭代训练过程.我们使用得到的粗略标签训练修改的 deep-SVDD

算法.由于标签较为粗糙,损失函数项的超参数 u 需要设置为较大值.修改的 deep-SVDD 算法的训练过程与原始 deep-SVDD 算法一样,首先固定超球面半径 R,使用随机梯度下降法优化网络权重 $\boldsymbol{W}$,然后再固定 $\boldsymbol{W}$,使用线性搜索优化 R,这一过程反复进行多次,直到收敛为止.修改的 deep-SVDD 算法在训练之后,就可以为训练数据计算新的概率标签,而这些新的概率标签与先前计算的粗略概率标签融合,就可以得到具有更准确概率标签训练数据集(如 1.3 节所述).

接下来,我们使用这些更准确的概率标签,再次训练修改的 deep-SVDD 算法.由于标签比之前更准确,损失函数项的超参数 u 需要适当调小.经过训练的 deep-SVDD 算法再次为训练数据集计算概率标签,并将结果与先前计算的概率标签进行融合.这个过程持续进行多次,用于训练的概率标签会越来越准确,deep-SVDD 算法也会被训练得越来越充分,进而算法的预测准确度也会越来越高.

1.6 框架参数设置

对于 autoencoder,我们为编码设置了 3 个隐藏层,分别包含 128,74,32 个节点,为解码设置了 3 个隐藏层,分别包含 32,74,128 个节点,并让它以 30 期(epoch),256 批次(batch)进行训练.deep-SVDD 算法具有 10 层的 CNN,每层包含 128 个节点,以 4 期 256 批次反复迭代训练,学习率为 0.000 5.

2 实验设置和结果

2.1 实验设置

为了评测本项目所提出框架的性能,我们首先将整体的事故数据集分为 2 部分:训练数据集和测试数据集.训练数据集的数据量占整个数据集的 80%,而测试数据集是舍弃额外正样本数据而剩下的正负样本平衡数据,占整个数据集的 20%.首先,我们测试了不同参数下该框架的性能表现,包括网络层数(1~20)、迭代次数(1~8)和学习率(0.000 01~0.001)参数.该实验可以验证 1.6 节所述参数设置的可行性.此外,我们还测试了不同迭代次数下(1~8)训练数据标签生成结果的准确性.该实验可以证明该框架可以得到准确的训练数据标签,进而得到高质量的训练结果.此外,该框架是概率级别的无监督学习框架,但是在 0/1 级别上使用了原始的 0/1 标签是监督学习框架,并且训练数据集的正负样本不平衡,所以我们将该框架的性能与监督的不平衡二分类算法 CCNN(cost-sensitive CNN)和 CDNN(cost-sensitive DNN)进行了比较[26].同时,该框架对原始的 deep-SVDD 算法做了修改,所以我们也和原始 deep-SVDD 算法进行了比较.为得到概率级别的结果,CCNN 算法和 CDNN 算法直接使用输出层的分类概率,而 deep-SVDD 算法计算样本到超球面中心的余弦距离并进行归一化.该实验可以证明该框架优于现有算法.实验在具有 4 个 GPU 节点、67 000 个 CUDA 内核和 128 GB 内存的高性能计算集群上部署进行,所有算法都使用 Python 和 Tensorflow 1.12.0 深度学习框架实现.

2.2 评测指标

我们分别在 0/1 级别和概率级别对本项目所提出框架进行了评测.

对于 0/1 级别,我们将预测概率值<0.5 的值算作 0,≥0.5 的值算作 1,然后将它们与测试数据集中的 0/1 标签进行比较.因此,评测指标为:

1) 真阳性样本量(TP).TP 是正确预测的正样本数,即算法预测发生且实际也发生的事故数量.

2) 假阳性样本量(FP).FP 是错误预测的正样本数,即算法预测发生但实际未发生的事故数量.

3) 真阴性样本量(TN).TN 是正确预测的负样本数,即算法预测未发生且实际未发生的事故数量.

4) 假阴性样本量(FN).FN 是错误预测的负样本数,即算法预测未发生但实际发生的事故数量.

5) 召回率($Recall$).$Recall=TP/(TP+FN)$.

6) 精确率($Precision$).$Precision=TP/(TP+FP)$.

7) 准确率($Accuracy$).$Accuracy=(TP+TN)/(TP+FP+TN+FN)$.

8) F 值($F\text{-}score$).$F\text{-}score=2Recall\times Precision/(Recall+Precision)$.

对于概率级别的数值,我们应该将预测概率值(或生成的训练数据标签)与实际概率值进行比较,但是数据集中没有任何概率标签.因此,将该框架预测概率值(或生成的训练数据标签)与 k-means 算法、autoencoder 进行比较,相同结果越多,表明该框架的召回率和准确率越高,评测指标为:

1) I_K 是与 k-means 算法相同的结果数量;

2) I_A 是与 autoencoder 相同的结果数量;

3) I_B 是与 k-means 算法和 autoencoder 都相同的结果数量.

2.3 不同参数实验结果

2.3.1 不同隐藏层数的比较

表 1 给出了不同隐藏层数时,本项目所提出框架在 0/1 级别上的结果.当隐藏层数小于 10 时,该框架具有相对较小的召回率;而当数目增加到 10 以上时,该框架具有相对较高且稳定的召回率、精确率和准确率.

Table 1 0/1-Level Results with Various Numbers of Hidden Layers

表 1 不同数量隐藏层的 0/1 级别结果

隐藏层	*Recall*	*Precision*	*Accuracy*	*F-score*
1	0.862	0.713	0.809	0.808
5	0.887	0.709	0.812	0.814
10	0.933	0.730	0.821	0.836
15	0.947	0.732	0.819	0.839
20	0.941	0.732	0.822	0.840

表 2 给出了不同隐藏层数时,本项目所提出框架在概率级别上的结果.类似地,当隐藏层数小于 10 时,该框架具有相对较小的 I_K,I_A,I_B 值;而当数目增加到 10 及以上时,该框架具有较高且较平稳的 I_K,I_A,I_B 值.因此,如 1.6 节所述,隐藏层的数量为 10 是一个合理的参数选择.

Table 2 Probability-Level Results with Various Numbers of Hidden Layers

表 2 不同数量隐藏层的概率级别结果

隐藏层	I_K	I_A	I_B
1	3 123	2 100	199
5	6 735	7 476	302
10	9 441	12 743	5 334
15	9 532	12 005	5 144
20	9 423	12 801	5 256

2.3.2 不同迭代次数的比较

表 3 给出了在不同迭代次数下,本项目所提出框架在概率级别上的结果.在迭代次数低于 4 时,该框架具有相对较小的 I_K,I_A,I_B 值,而当迭代次数增加到 4 及以上时,该框架具有较高且较平稳的 I_K,I_A,I_B 值.因此,如 1.6 节所述,迭代次数为 4 是一个合理的参数选择.

因为在不同迭代次数下,本项目所提出框架在 0/1 级别上的结果与其他算法相比没有太大差异,所以没有直接给出.这表明在训练时,该框架在 0/1 级别收敛时,可能并没有在概率级别收敛,所以概率级别的研究是非常有必要的.

Table 3 Probability-Level Results with Various Numbers of Iterations

表 3 不同迭代次数的概率级别结果

迭代次数	I_K	I_A	I_B
1	3 022	2 317	202
2	6 045	2 299	214
4	9 644	13 016	5 193
6	9 502	12 887	5 068
8	9 522	12 988	5 197

2.3.3 不同学习率的比较

除了 0/1 级别和概率级别的结果外,图 4 给出了在不同学习率下,损失函数数值随着训练批次增加的变化趋势.当学习率低于 0.000 5 时,损失函数值下降缓慢;当学习率增加到 0.000 5 以上后,损失函数值下降迅速,但是较高的学习率会导致损失函数值的频繁波动.因此,如 1.6 节所述,学习率为 0.000 5 是一个可行的相对最优参数.

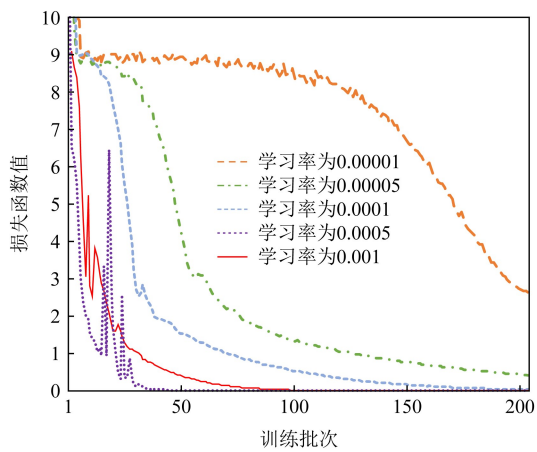


Fig. 4 Decrease of loss with various learning rates

图 4 不同学习率的损失值下降情况

2.3.4 训练数据标签生成结果

不同迭代次数下训练数据的标签生成结果与 2.3.2 节相似.在迭代次数低于 4 时,该框架具有相对较小的 I_K,I_A,I_B 值;而当迭代次数增加到 4 及以上时,该框架具有较高且较平稳的 I_K,I_A,I_B 值.第 8 次循环后,I_K,I_A,I_B 值分别可达 9 522,12 988,5 197.因此,该框架可以得到准确的训练数据标签,进而得到高质量的训练结果.

2.4 算法比较实验结果

表 4 给出了本项目所提出框架与 CDNN 算法、CCNN 算法和原始 deep-SVDD 算法的 0/1 级别的结果.这些方法都具有较高的真阳性样本量,但是相比 CDNN 算法,CCNN 算法、deep-SVDD 算法和该框架具有更高的真阴性样本量和更低的假阳性样本量.除此之外,这些方法都具有相近的召回率,但是 CCNN 算法、deep-SVDD 算法和该框架都具有更高的精确率、准确率和 F 值.图 5 给出了 CDNN 算法、CCNN 算法、原始的 deep-SVDD 算法和本项目所提出框架的 ROC 曲线.CCNN 算法、deep-SVDD 算法和该框架的 AUC 值高于 CDNN 算法.

Table 4 0/1-Lvel Comparison Results to Different Methods

表 4 不同方法的 0/1 级别结果

算法	*TP*	*TN*	*FP*	*FN*	*Recall*	*Precision*	*Accuracy*	*F-score*
CDNN	28 458	16 928	17 072	1 542	0.949	0.625	0.709	0.664
CCNN	26 586	22 139	11 861	3 414	0.886	0.691	0.761	0.724
deep-SVDD	28 225	22 369	11 631	1 775	0.941	0.708	0.790	0.808
本文框架	29 368	23 144	10 856	632	0.979	0.730	0.821	0.836

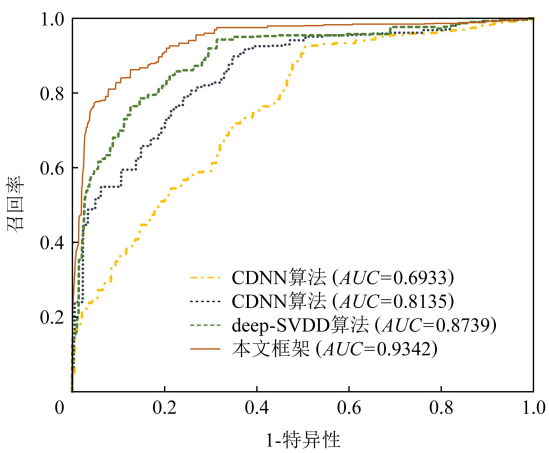


Fig. 5 ROC curves of the different methods

图 5 不同方法的 ROC 曲线

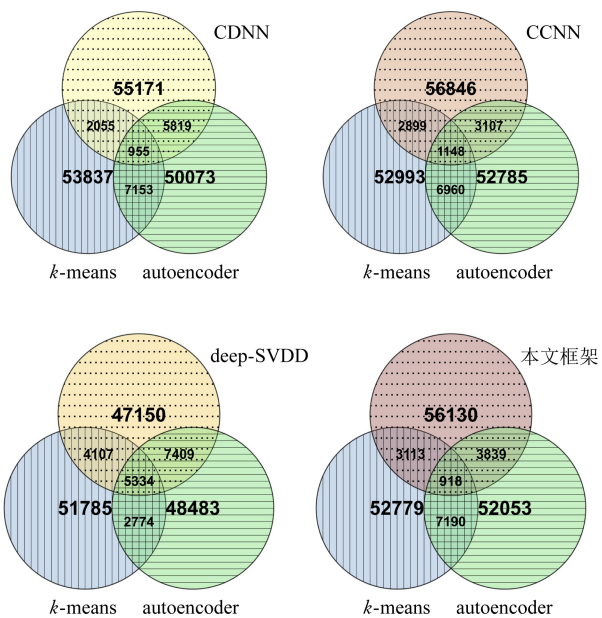


Fig. 6 Venn diagrams of prediction results among the different algorithms

图 6 不同算法的预测结果韦恩图

图 6 用韦恩图给出了本文所提出框架与 CDNN 算法、CCNN 算法和原始 deep-SVDD 算法概率级别的结果.与其他算法相比,该框架与 k-means 算法和 autoencoder 都具有更多相同的概率值.也就是说,该框架可以计算出更准确的概率值.特别地,尽管 CCNN 算法、deep-SVDD 算法和该框架在 0/1 类级别上的性能相似,但是该框架在概率级别上要优于 CCNN 算法和 deep-SVDD 算法.

3 总结与展望

本文首先通过融合 OSU,FARS 数据集,SHRP2 数据集分布相同的 Sim-SHRP2,以及自己驾驶汽车,得到了一个完整的车辆事故数据集.此外,我们还设计了一个概率级别的无监督深度学习框架来预测事故发生的准确概率值.该框架使用迭代的方式为数据集生成准确的概率标签,并使用这些概率标签来进行训练.我们分别在事故发生的 0/1 级别和概率级别上对该框架进行了评测,实验结果表明,该框架可以使用所得到的数据集来灵敏而准确地预测车辆事故.以该框架为基础,可以构建为车辆驾驶员提供不同级别事故预警的软件.

之后,我们将把这项工作与目标检测的识别方法相结合,使得车内设备可以自动、实时地获取交通标志[27]、地理情况、环境情况、交通情况和驾驶员行为等特征,进而构建最终的事故预警软件.

参 考 文 献

[1] National Highway Traffic Safety Administration. 2015 motor vehicle crashes: Overview [M] //Traffic Safety Facts Research Note. Washington, DC: NHTSA, 2016: 1-9

[2] Yang Kui. Deep learning for real-time crash prediction on urban expressways [C] //Proc of the 97th Transportation Research Board Annual Meeting. Piscataway, NJ: IEEE, 2018: No.18-04829

[3] Yuan Zhuoning, Xun Zhou, Yang Tianbao. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 984-922

[4] Lu Wenqi, Luo Dongyu, Yan Menghua. A model of traffic accident prediction based on convolutional neural network [C] //Proc of the 2nd IEEE Int Conf on Intelligent Transportation Engineering (ICITE). Piscataway, NJ: IEEE, 2017: 198-202

[5] Gite S, Himanshu A. Early prediction of driver's action using deep neural networks [J]. International Journal of Information Retrieval Research, 2019, 9(2): 11-27

[6] Liu Miaomiao, Chen Yongsheng. Predicting real-time crash risk for urban expressways in China [J]. Mathematical Problems in Engineering, 2017, 2017: 1-10

[7] Dixon R, Carl E, Chris F. Supervised machine learning for modeling human recognition of vehicle-driving situations [C] //Proc of 2005 IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2005: 604-609

[8] Jabon M. Facial expression analysis for predicting unsafe driving behavior [C] //Proc of IEEE Pervasive Computing. Piscataway, NJ: IEEE, 2010: 84-95

[9] Tambouratzis T. Combining probabilistic neural networks and decision trees for maximally accurate and efficient accident prediction [C] //Proc of the 2010 Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2010: 1-8

[10] Wahab A. Driving profile modeling and recognition based on soft computing approach [J]. IEEE Transactions on Neural Networks, 2009, 20(4): 563-582

[11] Lafferty J, Andrew M, Fernando C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] //Proc of the 18th Int Conf on Machine Learning. New York: ACM, 2001: 282-289

[12] Akin D, Bulent A. A neural network(NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics [C] //Proc of Scientific Research and Essays. Piscataway, NJ: IEEE, 2010: 2837-2847

[13] Lu Yuejing. Research on accident prediction of intersection and identification method of prominent accident form based on back propagation neural network [C] //Proc of 2010 Int Conf on Computer Application and System Modeling (ICCASM 2010), Vol1. Piscataway, NJ: IEEE, 2010: V1-434

[14] Rezaie M, Afandizadeh S, Ziyadi M. Prediction of accident severity using artificial neural networks [J]. International Journal of Civil Engineering, 2011, 9(1): 41-48

[15] Abdel A, Mohamed N, Uddin N, et al. Predicting freeway crashes from loop detector data by matched case-control logistic regression [J]. Transportation Research Record, 2004, 1897: 88-95

[16] Zheng Zuduo, Ahn S, Monsere C M. Impact of traffic oscillations on freeway crash occurrences [J]. Accident Analysis & Prevention, 2010, 42(2): 626-636

[17] Xu Chengcheng. Evaluation of the impacts of traffic states on crash risks on freeways [J]. Accident Analysis & Prevention, 2012, 47(1): 162-171

[18] Abdel A, Mohamed A. Real-time prediction of visibility related crashes [J]. Transportation Research Part C: Emerging Technologies, 2012, 24: 288-298

[19] Chen Chao. SDCAE: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data [C] //Proc of the 6th Int Conf on Advanced Cloud and Big Data(CBD). Piscataway, NJ: IEEE, 2018: 328-333

[20] Fridman L. Human-centered autonomous vehicle systems: Principles of effective shared autonomy [J]. arXiv preprint, arXiv:1810.01835, 2018

[21] Moosavi S. A countrywide traffic accident dataset [J]. arXiv preprint, arXiv:1906.05409, 2019

[22] Moosavi S. Accident risk prediction based on heterogeneous sparse data: New dataset and insights [C] //Proc of the 27th ACM SIGSPATIAL Int Conf Advances in Geographic Information Systems. New York: ACM, 2019: 33-42

[23] Zhang Dingwen, Han Junwei, Zhang Yu. Supervision by fusion: Towards unsupervised learning of deep salient object detector [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 4048-4056

[24] Ruff L. Deep one-class classification [C] //Proc of the IEEE Int Conf on Machine Learning. Piscataway, NJ: IEEE, 2018: 4393-4402

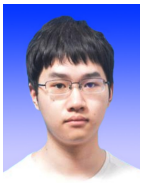
[25] An J, Sungzoon C. Variational autoencoder based anomaly detection using reconstruction probability [J]. Special Lecture on IE, 2015, 2(1): 1-18

[26] Krishnaveni C, Rani T. On the classification of imbalanced datasets [J]. International Journal of Computer Applications, 2012, 44(8): 1-7

[27] Li Xudong, Zhang Jianming, Xie Zhipeng, et al. A fast traffic sign detection algorithm based on three-scale nested residual structures [J]. Journal of Computer Research and Development, 2020, 57(5): 1022-1036 (in Chinese)
(李旭东, 张建明, 谢志鹏, 等. 基于三尺度嵌套残差结构的交通标志快速检测算法[J]. 计算机研究与发展, 2020, 57(5): 1022-1036)

Zhang Litian, born in 1998. Master candidate. His main research interests include big data based algorithm and artificial intelligence application.

张力天,1998 年生.硕士研究生.主要研究方向为基于大数据算法和人工智能应用.



Kong Jiayi, born in 1997. Master candidate. His main research interests include big data processing and machine learning.

孔嘉漪,1997 年生.硕士研究生.主要研究方向为大数据处理与机器学习.

Fan Yihang, born in 1996. Master candidate. His main research interests include machine learning and artificial intelligence application.

樊一航,1996 年生.硕士研究生.主要研究方向为机器学习与人工智能应用.

Fan Lingjun, born in 1982. PhD, assistant professor. Member of CCF. His main research interest is developmental tactic of information technologies, including intelligent city, big data, blockchain and industrial Internet.

范灵俊,1982 年生.博士,助理研究员,CCF 会员.主要研究方向为信息技术发展战略,包括智慧城市、大数据、区块链、工业互联网等.

Bao Ergude, born in 1984. PhD, associate professor and PhD supervisor. Member of CCF. His main research interest is big data based algorithm and artificial intelligence application, and the areas include transportation and biomedicine.

包尔固德,1984 年生.博士,副教授,博士生导师,CCF 会员.主要研究方向为基于大数据的算法和人工智能应用,应用领域主要包括交通运输和生物医疗.