Triple-O for SHL Recognition Challenge: An Ensemble Framework for Multi-class Imbalance and Training-testing Distribution Inconsistency by OvO Binarization with Confidence Weight of One-class Classification

Jinhua Su*
2017201620@ruc.edu.cn
Renmin University of China
China, Beijing

Yuanyuan Zhang
yuanyuan.zhang@haochezhu.club
Beijing Baixingkefu Network Technology Co., Ltd.
China, Beijing

ABSTRACT

SHL Recognition Challenge provides phone sensor data for recognizing eight modes of locomotion and transportation(activities). Most motion sensor data is user-dependent, so the model's generalization ability requires various data at the user level. The SHL dataset is collected only by three users, and category imbalance and distribution gap between training data and testing data can be significant obstacles for this challenge. In comparison to exploring the large and deep model structure, some model-free tricks play a more critical role in the previous challenges, such as transfer learning, re-sampling and anti-overfit trick. SHL challenge also notices that problem and provides user-independent sensor data in 2021. This paper analyses the data and finds out that category imbalance and distribution inconsistency are still the obstacles. This paper focuses on evaluating different methods to improve the general predicting ability at the setting of category imbalance and training-testing distribution inconsistency. Besides, this paper puts forward a new ensemble framework called triple-O, using OvO binarization and one-class classification. The results show that the OvO binarization ensemble gets better results on the hard-to-distinguish class than re-sampling and re-weighting. One-class classification can be an anomaly detection to re-weight the meta learners to tackle the distribution gap. Triple-O can be plug-and-play and pave the way for exploring complex model structures. This paper introduces the solutions from the team GoodGoodDriveDayDayUp.

CCS CONCEPTS

• **Computing methodologies → Bagging**.

KEYWORDS

ensemble, multi-class classification, one-class classification, activity recognition

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09…$15.00
https://doi.org/10.1145/3460418.3479375

ACM Reference Format:
Jinhua Su and Yuanyuan Zhang. 2021. Triple-O for SHL Recognition Challenge: An Ensemble Framework for Multi-class Imbalance and Training-testing Distribution Inconsistency by OvO Binarization with Confidence Weight of One-class Classification. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3460418.3479375

1 INTRODUCTION

SHL(short for Sussex-Huawei Locomotion) Recognition Challenge provides a phone sensor dataset for recognizing eight modes of locomotion and transportation activities, attracting worldwide researchers since 2018[16, 17, 20]. Activities Locomotion can be regarded as a multi-class sequence labelling task whose target is a multi-class predicting label vector given a variable-length sequence. There are two mainstream solutions; one is processing features with sliding window and applying machine learning classifier. The other uses deep sequence labelling models like RNN, LSTM and Transformer to give predictions one by one. The former depends on feature engineering, and the latter extracts the feature automatically but sometimes fails to capture the long-distance dependency, which may be necessary for recognizing track transportation. For the most time, the deep learning models have better performance.

A survey on the overview paper of the SHL Recognition Challenge 2018, 2019 and 2020 gives several notable and confusing conclusions exceeding the expectations.

- There is no apparent advantage of the deep and heavy model compared to some naive but robust machine learning methods.
- With re-sampling or other data imbalance methods, models still give the worst prediction for the minor category.
- Overfitting is common; most feeling-good models get a lower ranking. The explicitly underfit model won first place.
- Traning only on the training dataset makes many participants doubt life. Finetuning on the validation dataset works.

Suddenly turn back to the dataset, where the principal contradiction lay, and find out category imbalance and distribution gap between the training and validate dataset. The whole dataset is collected by only three voluteers. The dataset is divided into training, validation and testing as the following.

- Training data is collected by User A in 59 days.
- Validation data is collected by User B and User C in 4 days.

- Testing data is collected by User B and User C in 39 days.

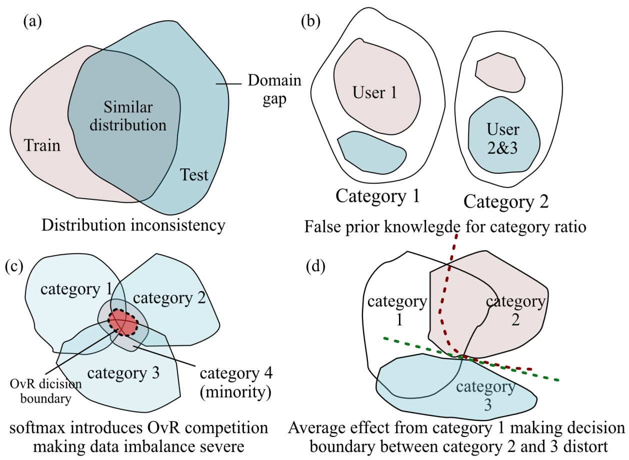

Figure 1: Distribution insistency and two problems introduced by multi-class softmax function
Notes: Multi-modality and user-dependent difference make softmax loss work worse for multi-class data imbalance. Subfigure (a) and (b) are venn diagrams for distribution consistency in user-level and category-level. Subfigure (c) and (d) are venn diagrams for OvR imbalance and average effect causing low intra-class separability from softmax loss.

The distribution inconsistency in user-level and category-level arise a domain adaptation problem for challenge participants. The concept map is shown in fig.1. Different user has different phone-using manners at the same activity. Furthermore, different locomotion and transportation activities depend on the real-world transportation conditions and user commuting habits, which indicates that the category ratio of one user may vary during weekdays and weekends. Multi-modal distribution in different levels makes recognition challenging. Facing the same problem, face recognition use centre loss and marge loss to realize domain adaptation. Lack of user category labels, it is hard to apply centre loss for intra-class compactness directly. However, inter-class separability can be integrated for the activities recognition task. This paper is aimed to alleviate the above problems drawing lessons from the research of face recognition[10, 11, 15, 21].

$$p_k = \frac{exp(z_k)}{\sum_{i=1}^{K} exp(z_i)} \tag{1}$$

The softmax function, shown in 1, should be to blame for both low inter-class separability and minor-class neglecting at the setting of multi-class classification[6]. Indeed, A flow of researches modifying the softmax to improve face recognition. The softmax gives relatively bad predictions on the minority class, causing a low macro-F1, one of the most important metrics. Most time, we view softmax as a multi-class extension of the logistic for binary classification. [8, 9, 12] all shows that softmax treats all class labels equally and poses a competition between true and other class labels. That property causes two problems. One is the so-called data imbalance problem. Softmax loss makes the model sensitive to the category prior ratio, which causes minor classes to rarely get a big probability prediction as the category number becomes large. The other is the multi-modality neglect problem; from the perspective of ensemble learning, the softmax can be regarded as the OvR(one v.s. rest) approach, where the multi-modality of the rest may be harmful to model performance. The negative class of the OvR approach consists of all the remaining non-target classes. Softmax treats all non-target class that shares non-trivial similarity with the target class, contributing to the so-called average effect. The average effect of multi-modality leads to low intra-class separability. Looking at the confusion matrix of previous SHL Challenge results, similar activities sharing non-trivial similarities are hard to distinguish, such as car and bus.

Splitting the multi-class classification task into several OvO(one v.s. one) binary-class classification gives a solution to low intra-class separability. Indeed, [3] has done the experiments on the relatively simple task for this proposal. The results show OvO methods did outperform the OvR methods, especially for the imbalanced situation. From the perspective of curriculum learning, binarization can also fasten model learning procedures given the same optimizer. Nowadays, more classification models are modified for sequence labelling, claiming a better result than classical state transfer models, such as Hidden Markov Model. Since, in most cases, the state space is discretization, they are all sequence multi-class classification. It will be a new perspective to discuss the imbalance of the states.

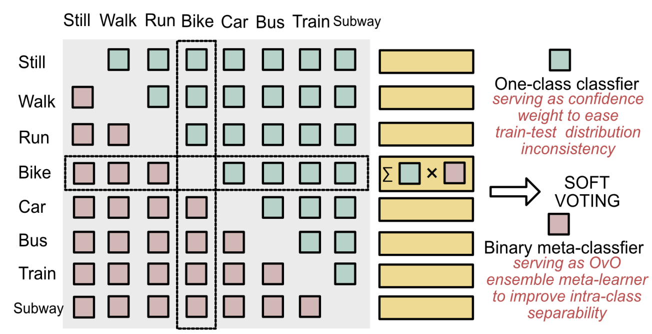

Figure 2: Pipeline of Triple-O ensemble for SHL Recognition Challenge
Notes: Eight-class classification are divided into 28 binary classification. For each binary learner, corresponding one-class classifier(sometimes viewed as the outlier detector) is trained to serve as confidence weight for soft voting.

This paper put forward a ensemble framework, called Triple-O, designed to tackle with multi-class imbalance and distribution inconsistency in the SHL Recognition Challenge. The pipeline flowchart is shown in fig. 2. Triple-O applies the OvO method to tackle the multi-class imbalance problem caused by the softmax method. Experiments show its efficiency compared to other data imbalance solutions, like re-sampling and re-weighting. What is more, another challenge of this dataset is the training and testing distribution inconsistency. To tackle this problem, we use a one-class classification method[1, 14] to detect the outlier from the training distribution. The predicting confidence can be used as the weight of meta-learner for soft voting. This ensemble uses the cascade framework, which is creative and needs to be further researched. The OvO ensemble plus One-class classification weighting pipeline is named Triple-O. Triple-O can be plug-and-play and pave the way for exploring complex model structures. To our best knowledge, this method has not been further explored. The most relevant work we can find is [2].

2 DATASET

2.1 Description

The SHL Challenge has been held every year since 2018. The goal for 2021 is to recognize eight modes of locomotion and transportation(activities) in a user-independent manner based on radio data, including GPS reception, GPS location, WiFi reception and GSM cell tower scans. The raw radio data contains a list of different signals data. The number of the signals are uncertain, but it may be the same in the same area, such as WiFi or GSM data. Location data is processed from GPS data. The data are collected in about 1 Hz frequency. The descriptive information and signal subtypes are described in the table 1.

2.2 Preprocessing for classification

To deeply evaluate the multi-class data imbalance, other data property, which may influence the model performance, requires adjustment. The processing procedure is shown at the tab.2. The raw data need to be alignment by the timestamp. Since the frequency of all data is close to 1 Hz, we drop to save the integral part of the timestamp and group the data by timestamp. This procedure may drop a few rows of highly missing data. Location is derivated twice to get the speed and the acceleration. The one-hot representation is used to encode each GPS satellite signal. We use PCA(principal component analysis) to get the top fifty valuable components to reduce the dimension. On the other hand, a list of WiFi and GSM Cells signals on each timestamp is aggregated by the mean, max, min, std operation. Missing information sometimes will indicate the low signal environment, which can help the model recognize some transportation vehicles. A bool variable is built for each column. The 2 minutes length window centred on the target timestamp is used to calculate the quantile information. Using quantile statistics of the window data avoid overfitting and postprocessing, for the quantile information will not change a lot in a short time.

3 APPROACHES

This paper tries to apply binarization strategy on the typical sequence multi-class classification task, activity recognition task. SHL Recognition Challenge[4, 19] provides an open dataset with complete mobile sensor data, which is a close-to-life case. Besides the imbalanced multi-class classification, Another challenge of this dataset is the training and testing distribution inconsistency.

The background of this work is directly using softmax-based multi-class classifier make models bad at general predicting ability, especially on the imbalanced(minority) class.

3.1 Why do models with softmax loss degrade at the setting of multi-class imbalance

Multi-class classification is deeply associated with Binary-class Classification but is more challenging[7]. The problem caused by data imbalance exists nearly everywhere. Contract to binary-class classification, data imbalance becomes more complex. For most learners, multi-class prediction is a combination of OvR(one v.s. rest) [22] predicting score. The probability for each class is produced by the softmax method, which can be taken as the soft voting machine. The relative score determines the final decision of the machine. For the most time, the absolute value of the softmax probability tends to be close to 0 or 1.

Why the more classes there are, the worse prediction for the minority. OvR predicting score and softmax method should blame. The more classes there are, the more extreme for an OvR learner to predict. With expositional function, the predicting probability for minority class may be close to 0. So sharpening or darkening operation will be applied before the softmax procedure to make the absolute value of probability more reasonable. However, the hyperparameter of sharpening or darkening is hard to set.

It is hard to modify the predicting score. So the specific solutions are to improve the influence of the minority during the training procedure. Re-sampling and Re-weighting have a similar effect on the training loss. But unlike binary-class classification, multi-class classification unavoidably gives an imbalanced input for OvR training.

The flowchart of 3 shows common approaches and our approach. Directly training a softmax-based model on the raw data prefer to split the smaller decision space, which means the model will give a more negligible probability for the minority class. Re-Sampling(or over-Sampling) and re-weighting finally enlarge the decision space, making the prediction on the minority class more reliable. With the number of multi-class increase, the decision space for each class is more petite. The hyperplane of the multi-class will be more complicated and then cause the generalized ability constrained. Dividing the multi-class problem into several binary problems can make the hyperplane between two classes simpler and more robust. The shortcut is its generalization may cause false-positive cases since it only uses partial data of the whole multi-class dataset at once. Using One-class Classification to find a suitable boundary for controlling the generalization ability is an excellent idea for weighting the meta-learner confidence. Binarization plus One-class classification may be an interesting pipeline for multi-class classification problems.

3.2 Common solutions for data imbalance

The learning phase and the subsequent prediction of machine learning algorithms can be affected by the imbalanced data set problem. The balancing issue corresponds to the difference in the number of samples in the different classes. One of the easiest ways to fix it is to re-sampling. SMOTE (Synthetic Minority Over-sampling Technique)[5] is the most common method to improving the number of the minority class. The repeated sample indirectly make the loss on the minority larger. So another more direct and computing less way is to give minority loss part a large weight[13]. This

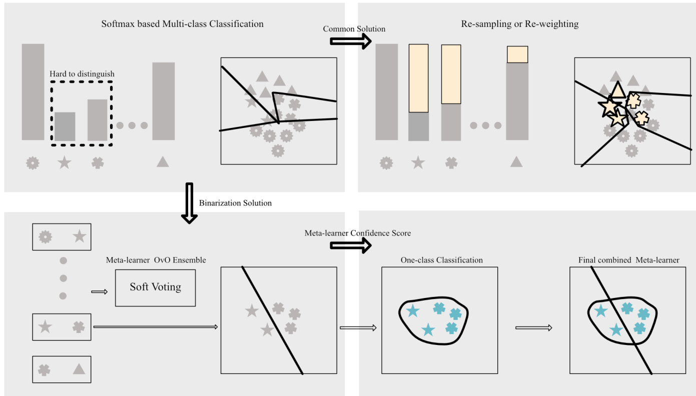

Figure 3: Contrast analysis of common re-balance methods and our solution
Notes: Our method is designed for multi-class imbalance and distribution inconsistency.

	size	columns	signal subtypes
Location	911109	Accuracy, Latitude, Longitude, Altitude	None
GPS	1322749	A list of satellite signals, each containing SNR, Azimuth, Elevation	About 200 different satellites
WiFi	1459351	A list of WiFi signals, each containing Frequency, Capabilities	None
GSM Cells	1324881	A list of towel signals, each containing signal strength and level	LTE, WCDMA and GSM

Table 1: The description of the initial training data.
Notes: Different sources of sensor data all are attached with the timestamps. The size means the counting of the timestamps.

Process Period	Locotion	GPS	Cell	Wifi	Details
Alignment	Yes	Yes	Yes	Yes	Align and bind based on label's timestamp
Missing Filling	Yes	Yes	Yes	Yes	Using mean value to fill the NaN and None
Manual Feature	Yes	No	No	No	Calculate 1-order and 2-order difference
PCA selecting	Yes	Yes	No	No	Apply PCA on one-hot encoding and select top-50
Variate-length	No	No	Yes	Yes	Use mean, standard variance and quantile to get Invariate-length feature
Sliding Window	Yes	Yes	Yes	Yes	Use 2-minute window to calculate the percentile feature for each timestamp as centre
Normalization	Yes	Yes	Yes	Yes	Use Z-score to normalize the feature for model input

Table 2: The data processing procedure form.

method, called re-weighting, can achieve the same result when the loss function is log-based, such as log-likelihood or cross-entropy.

F1(%)	1	2	3	4	5	6	7	8	Mean
Train-Te	96.3	92.5	91.1	97.1	91.2	91.1	92.5	95.0	93.3
Valid-Fu	58.7	61.2	0.0	52.9	60.6	15.0	54.1	71.9	46.8

Table 3: The result of random forest trained only on the training-training part.
Notes: Train-Te is short for Training-Testing. Valid-Fu represents Validation-Full.

Activity	Still-1	Walking=2	Run=3	Bike=4
Train	122301	122766	42276	117401
Valid	29836	26149	2772	12031
Activity	Car=5	Bus=6	Train=7	Subway=8
Train	158695	141643	156260	119183
Valid	20473	9178	21808	21709

Table 4: Frequency of different activities in the train and validation dataset
Notes: Valid is short for validation.

3.3 Binarization: OvO ensemble

Indeed a model trained on the imbalanced data can get a nice evaluation result with micro-F1. Macro-F1 is designed for a balanced prediction for each class. Model Ensemble and predicting by weighted voting seem to be another re-weighting method for improving the minority class. There are two binarization methods: OvR(one v.s. rest) and OvO(one v.s. one). The softmax-based method is a special OvR ensemble where all meta-learners share the same encoder structure and model parameters. That is to say,

F1(%) Valid Test	1	2	3	4	5	6	7	8	Mean
RF(direct)	93.5	91.8	67.4	89.4	87.3	71.2	92.7	91.7	85.6
RF(Re-Sample)	90.5	87.4	74.3	89.4	87.9	75.6	92.3	91.5	86.1
RF(Re-Weight)	89.2	88.4	73.9	90.1	86.8	77.2	91.8	92.3	86.2
RF(OvO)	97.6	92.8	84.7	99.9	94.1	85.2	99.3	98.3	94.0
RF(Tirple-O)	87.6	88.8	81.7	92.3	90.2	88.2	85.3	87.3	87.6
RF(OvO+Train)	96.4	91.5	80.3	99.9	92.7	81.1	98.3	97.4	92.2

Table 5: Evaluation on different learning strategies on validation dataset
Notes: RF is short for random forest.

	1	2	3	4	5	6	7	8
1	4114	88	0	6	0	0	56	0
2	29	4572	0	0	21	0	28	0
3	0	100	204	0	0	0	0	0
4	0	0	0	2082	0	0	0	0
5	92	30	0	0	3516	266	0	145
6	38	486	0	0	0	1784	40	0
7	0	0	0	0	0	0	3487	0
8	0	70	0	0	0	0	0	3999

Table 6: Random forest(OvO+Train) confusion matrix on Validation-Testing(column: predict, row: true)

the OvR ensemble will still suffer the same imbalance problem as multi-class classification. So simplifying the multi-class imbalance problem by OvO agrees with the theory of curriculum learning: "divide and conquer" or "easy first then complex". For each meta-learner, binary-class classification is easier to learn and solve the severity of the imbalance.

3.4 One-class classification

Furthermore, the risk of OvO ensemble lays out explicitly. Meta-learner works well when the actual label is in one of the binary-class. However, the prediction will be a nuisance if the actual label is neither one. Give the learner a confidence level help the OvO ensemble give a believable prediction. Judging the data is one of the binary-class or not is a One-class classification problem, which sometimes is similar to anomaly detection or outlier detection. The easiest method is to find the smallest hypersphere with all the binary-class data contained in it.

4 EXPERIMENTS & RESULTS

We focus on the imbalance problem and distribution gap; choose a robust baseline will give more reliable results instead of choosing an RNN-based encoder to tackle the sequence labelling iteratively. We use quantile information of the window data—we split each entire sequence into 20% testing part and 80% training part. Combine the partial sequences together to train a Random Forest model.

4.1 Evaluating model

We choose the Random Forest with 100 trees and 20 as the max depth for the multi-class classification. And 30 trees and depth up to 10 is tested to be the OvO ensemble to avoid the overfit. The node splitting method is the Gini coefficient. For the OvO ensemble, we choose OneVsOneClassifier from package sklearn. It constructs one classifier per pair of classes. At prediction time, the class which received the most votes is selected. In the event of a tie (among two classes with an equal number of votes), it selects the class with the highest aggregate classification confidence by summing over the pair-wise classification confidence levels computed by the underlying binary classifiers.

4.2 Device and time cost

This work focus on the ensemble pipeline framework to tackle multi-class imbalance. Our meta-learner, random forest, is not heavy-parameter. Most experiments are executed on a 8 core CPU, 16G Memory server. The final ensemble checkpoint takes up 142MB of space. Since it can be easily converted into a plug-and-play wrapper, meta-learner can be replaced by other deep learning models. The one-class classifier nearly takes no space. So the final checkpoint is believed to take 28 times larger space than the meta-learner. As for the Triple-O ensemble of the random forest, it takes 5 minutes to train and within 1 minute to evaluate the testing dataset.

4.3 Training on initial training dataset

The problem of distribution gap between training and testing has been mentioned in the previous challenge[17]. The main cause is the different patterns of the data collectors. The training dataset is all collected by person A, while the validation dataset and testing dataset are collected by B and C. The first thing we do is measure the gap. We split the training dataset into 80% training-training part and 20% training-test part and train the Random Forest model only on the training-training part. To measure the gap, we contrast the prediction of the training-testing part and the validation-full dataset. The results are shown in the table 3. The gap is large and Running, and Bus prediction has an awful result on the validation dataset. To dig out the specific reason, we counting the activity labels for the different datasets shown in the table 4. It is no surprise that the Run and the Bus are relatively smaller in the validation dataset. The further question is whether the B and C activities have the learnable pattern, which can be different from the A's. For predicting the B and C activities, directly learn the pattern from the validation is reasonable.

4.4 Training on rebuild training dataset

We split the validation dataset into 80% validation-training part and 20% validation-testing part and directly train the model on the validation-training and evaluated it on the validation-testing part. At this time, we focus on solving the data imbalance problem for the multi-class classification problem. The results of applying different learning strategies on the meta-learner are shown in the table 5. Directly training on the validation-training part still performed relatively worse in the prediction of Run and Bus. We are applying the re-sampling and re-weighting methods by package imlearn, an improvement on the Run and Bus. The F1 for other activities also drops, which is reasonable that the previous give a fairly good prediction for the majority. The mean of all activities F1 score(equals to macro-F1) do improve in a small amount. Using the binarization ensemble will enhance the total parameters; the final results appear to be overfitting. By looking up the training evaluation, 96.2 of macro-F1 indicates the overfitting is not severe.

To apply the most straightforward One-class Classification, we calculate the centre of binary data collections. Furthermore, select 90% quantile of the distance between data and centre as the radius r of the sphere. The above OvO voting procedure is to calculate the sum of probability of each binary meta-learner. We estimate 28 spheres(decision boundary for our one-class classifier) for the 28 meta-learner of the OvO ensemble and use the following confidence equation(shown in 2) to get a weighted probability for final prediction. The results show that every type of activity becomes worse and worse. It is hard to claim the failure of One-class classification for limited data and lack of hyperparameter tuning. If the radius is too small, one-class classification will limit the model generalization ability. On the other hand, whether the sphere is a suitable boundary for data has not been explored.

$$confidence = exp^{-\frac{(d-r)^2}{r^2}}), d > r \tag{2}$$

Finally, add the training-training dataset for the OvO ensemble pretraining to make the model more robust. The score drops a little. The confusion matrix of this OvO ensemble evaluated on the validation-testing is shown in the table 6.

5 CONCLUSION & DISSCUSSION

For SHL Recognition Challenge, the binarization strategy works on the setting of distribution inconsistency and class imbalance. The OvO ensemble method explicitly outperforms in contrast to the re-sampling and re-weight methods. To further tackle distribution inconsistency, we try the One-class classification to re-weight the contribution of each meta-learner to the final prediction. It is creative and needs to be further researched. The recognition result for the testing dataset will be presented in the summary paper of the challenge[18].

REFERENCES

[1] Md Nasim Adnan and Md Zahidul Islam. 2015. One-vs-all binarization technique in the context of random forest. In *Proceedings of the european symposium on artificial neural networks, computational intelligence and machine learning*. 385–390.

[2] Dong Chen, Xiao-Jun Wang, Changjun Zhou, and Bin Wang. 2019. The distance-based balancing ensemble method for data with a high imbalance ratio. *IEEE Access* 7 (2019), 68940–68956.

[3] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition* 44, 8 (2011), 1761–1776.

[4] Hristijan Gjoreski, Mathias Ciliberto, Lin Wang, Francisco Javier Ordonez Morales, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2018. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access* 6 (2018), 42592–42604.

[5] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*. Springer, 878–887.

[6] Rong Jin and Jian Zhang. 2007. Multi-class learning by smoothed boosting. *Machine learning* 67, 3 (2007), 207–227.

[7] Bartosz Krawczyk, Bridget T McInnes, and Alberto Cano. 2017. Sentiment classification from multi-class imbalanced twitter data using binarization. In *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 26–37.

[8] Xiangrui Li, Xin Li, Deng Pan, and Dongxiao Zhu. 2020. On the learning property of logistic and softmax losses for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4739–4746.

[9] Jun-Ying Liu and Bin-Bin Jia. 2020. Combining one-vs-one decomposition and instance-based learning for multi-class classification. *IEEE Access* 8 (2020), 197499–197507.

[10] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8679–8687.

[11] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. 2017. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507* (2017).

[12] Yazhou Ren, Peng Zhao, Yongpan Sheng, Dezhong Yao, and Zenglin Xu. 2017. Robust softmax regression for multi-class classification with self-paced learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2641–2647.

[13] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40, 12 (2007), 3358–3378.

[14] David Martinus Johannes Tax. 2002. One-class classification: Concept learning in the absence of counter-examples. (2002).

[15] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5265–5274.

[16] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2019. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2019. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 849–856.

[17] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2020. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 351–358.

[18] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2021. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2021. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*.

[19] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2019. Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset. *IEEE Access* 7 (2019), 10870–10891.

[20] Lin Wang, Hristijan Gjoreskia, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2018. Summary of the sussex-huawei locomotion-transportation recognition challenge. In *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers*. 1521–1530.

[21] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2019. A comprehensive study on center loss for deep face recognition. *International Journal of Computer Vision* 127, 6 (2019), 668–683.

[22] Wei Wu, Xiaorong Gao, and Shangkai Gao. 2006. One-versus-the-rest (OVR) algorithm: An extension of common spatial patterns (CSP) algorithm to multi-class case. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2387–2390.