symmetry

MDPI

Article

A Semi-Supervised Tri-CatBoost Method for Driving Style Recognition

Weirong Liu [1], Kunyuan Deng [2], Xiaoyong Zhang [1,*], Yijun Cheng [2], Zhiyong Zheng [2], Fu Jiang [1] and Jun Peng [1]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China; frat@csu.edu.cn (W.L.); jiangfu0912@csu.edu.cn (F.J.); pengj@csu.edu.cn (J.P.)
[2] School of Automation, Central South University, Changsha 410083, China; dengky@csu.edu.cn (K.D.); yijuncheng@csu.edu.cn (Y.C.); zhiyongzheng@csu.edu.cn (Z.Z.)
* Correspondence: zhangxy@csu.edu.cn; Tel.: +86-0731-8253-9616

Received: 23 January 2020; Accepted: 21 February 2020; Published: 26 February 2020

Abstract: Driving style recognition plays a key role in ensuring driving safety and improving vehicle traffic efficiency. With the development of sensing technology, data-driven methods are more widely uesd to recognize driving style. However, adequately labeling data is difficult for supervised learning methods, while the classification accuracy is not sufficiently approved for unsupervised learning methods. This paper proposes a new driving style recognition method based on Tri-CatBoost, which takes CatBoost as base classifier and effectively utilizes the semi-supervised learning mechanism to reduce the dependency on data labels and improve the recognition ability. First, statistical features were extracted from the velocity, acceleration and jerk signals to fully characterize the driving style. The kernel principal component analysis was used to perform nonlinear feature dimension reduction to eliminate feature coupling. CatBoost is an ensemble of symmetric decision trees whose symmetry structure endows it fewer parameters, faster training and testing, and a higher accuracy. Then, a Tri-Training strategy is employed to integrate the base CatBoost classifiers and fully exploit the unlabeled data to generate pseudo-labels, by which the base CatBoost classifiers are optimized. To verify the effectiveness of the proposed method, a large number of experiments are performed on the UAH DriveSet. When the labeling ratio is 50%, the macro precision of Tri-CatBoost is 0.721, which is 15.7% higher than that of unsupervised K-means, 1.6% higher than that of supervised GBDT, 3.7% higher than that of Self-Training, 0.7% higher than that of Co-training, 1.5% higher than that of random forest, 6.7% higher than that of decision tree, and 4.0% higher than that of multilayer perceptron. The macro recall of Tri-CatBoost is 0.744, which is also higher than other methods. The experimental results fully demonstrate the superiority of this work in reducing label dependency and improving recognition performance, which indicates that the proposed method has broad application prospects.

Keywords: driving style recognition; semi-supervised learning; label dependency; tri-training; CatBoost

1. Introduction

A driver's driving style is an important factor that needs to be considered in energy management systems and advanced driver assistance systems of electric vehicles [1,2]. It can be understood as a driving tendency, which influences the driver to complete driving tasks [3,4]. The recognition of driving style is helpful to understand driving behaviors. And it is an essential condition to enhance traffic safety, promote energy efficiency, and improve driving comfort. Therefore, the accurate recognition of driving style is attracting extensive research interests [5,6].

The existing driving style recognition methods are mainly divided into three types: the rule-based method, the model-based method and the data-driven method [1]. The rule-based method classifies the driving style according to the predefined thresholds [7]. The model-based approach describes the driving style by establishing an appropriate mathematical model [8]. However, rule-based methods mainly rely on technical knowledge and expertise, which may be inflexible and incomplete. On the other hand, mathematical models are complicated and difficult to establish due to the complexity and uncertainty of driving behaviors. In contrast, the data-driven method derives the style recognition model directly from the historical driving data using statistical and machine learning methods. This type of method has the advantages of high accuracy, low complexity, and strong generalization ability [9–11]. Therefore, a large number of data-driven methods have been applied to the recognition of driving style.

As one of the most important data-driven methods, the supervised learning method has achieved good results in the field of driving style recognition. Wang et al. developed an efficient driving style recognition method based on the support vector machine, which classifies drivers as aggressive and moderate [12]. Brombacher et al. used an artificial neural network to calculate aggressive scores for driving styles. According to the detected driving events, the overall driving style is divided into five types [13]. Xie et al. utilized a random forest as the style classification model for maneuver-based driving behaviors [14]. Sun et al. established a model based on the multi-dimension gaussian hidden Markov process to achieve accurate and reliable driving style recognition [15]. Although the supervised learning method has a satisfactory classification performance, a mass of labeled data is indispensable. If the labeled data are limited, the trained classifier cannot have strong classification ability. Moreover, it is difficult to provide high-quality labels for a large amount of data, which requires a lot of time and effort [16]. Therefore, for supervised learning methods, the reliance on the labeled data severely limits their practical application in the identification of driving style.

To avoid the problem of data labeling, some researchers have attempted to carry out unsupervised algorithms on unlabeled data for driving style recognition. Guo et al. proposed an unsupervised deep learning method. The method uses an autoencoder for feature learning and then performs driving style clustering recognition through a self-organizing network [17]. Ozgul et al. evaluated the performance of various unsupervised clustering methods on driving style recognition, such as k-means, spectral clustering, balanced iterative reducing and clustering using hierarchies [18]. Feng et al. used the support vector clustering to implement a driving style classification method based on driving events [19]. The essence of these methods is to discover the underlying structure of different driving styles from collected data directly, without manual intervention. The main advantage of unsupervised learning is that the classification does not require labeling data. Moreover, unsupervised learning is easy to deploy, and does not rely on a priori experience. However, the classification performance of unsupervised learning methods is not accurate enough. For boundary data between two styles, the classification difficulty is particularly significant. Therefore, the recognition ability of unsupervised learning methods makes it difficult to meet the requirements of practical driving style recognition.

Based on the above analysis, it is of great significance to develop a driving style recognition method that not only reduces the dependence on labels but also has a good performance. Compared with other semi-supervised strategies, Tri-Training can use the disagreement between three base classifiers to mine information contained in unlabeled data more effectively. Thus, Tri-Training [20] is very suitable for driving style recognition under limited labels. For Tri-Training, the selection of base classifier is of great importance. Categorical Boosting (CatBoost) [21] has an excellent classification accuracy and generalization ability, and it is the primary choice of base classifier comparing with other data-driven methods.

To solve the above challenges, this paper proposes an effective semi-supervised driving style recognition method. First, a large number of features are extracted from velocity data using statistical methods to characterize driving styles. Then, the kernel principal component analysis is used for nonlinear dimensionality reduction to eliminate the feature redundancy. Finally, Tri-CatBoost is

proposed to identify driving styles, combining the advantages of the semi-supervised learning strategy Tri-Training and the ensemble learning classification algorithm CatBoost. In this method, the unlabeled data are used to extend labeled data to optimize three CatBoost base classifiers through the Tri-Training strategy. Moreover, each base classifier plays an equivalent role and is further fused into a strong classifier. By learning the potential information contained in unlabeled data, the proposed method can reduce the demand for labeled data and effectively identify driving styles. The contributions of this paper are as follows:

- The semi-supervised learning strategy, Tri-Training, is applied to driving style recognition for the first time. It can fully mine the information of unlabeled data to reduce the label dependency and help to train a better recognition model. The introduction of the concept of semi-supervised learning provides a new research idea for driving style recognition.
- The CatBoost algorithm is performed as the base classifier for driving style recognition. As an improved ensemble learning algorithm, it has powerful classification and generalization capabilities. It can ensure accurate estimation of driving styles and is of great significance for accelerating practical applications of autonomous driving in intelligent transportation.
- A large number of comparisons with different supervised learning methods, semi-supervised strategies, and base classifiers are performed. The experimental results verify that the proposed method can improve the classification precision and recall compared to the existing methods when labeling data is limited. The comparison provides an effective evaluation on using semi-supervised for classic classification issues.

The rest of this paper is organized as follows. Section 2 demonstrates the main mechanism of the proposed method. Section 3 introduces the experimental data and feature processing. Section 4 validates the proposed method through lots of experiments and analyzes the results. Section 5 discusses the main findings of the study. Section 6 draws a conclusion.

2. The Proposed Method

2.1. Framework

The framework of the proposed semi-supervised driving style recognition method is shown in Figure 1. The proposed method uses a large number of unlabeled data to assist labeled data to improve learning performance. The method includes two stages: the offline stage and the online stage. In the offline stage, after data segmentation, feature extraction is used to form multiple indicators to represent the driving style. Kernel Principal Component Analysis (KPCA) is used to realize nonlinear dimensionality reduction, and standardization is used to accelerate convergence. Then, the driving style recognition model is built through the Tri-Training strategy. Specifically, three CatBoost classifiers are initialized with labeled data and are used to determine the pseudo-labels of unlabeled data. The pseudo-labeled data and the raw labeled data are put together to retrain classifiers. The three classifiers are iteratively updated and are finally ensemble into the semi-supervised recognition model Tri-CatBoost. In the online stage, the online data are processed in the same way as in the training stage, and are then provided for the trained Tri-CatBoost model to obtain the result of driving style recognition.

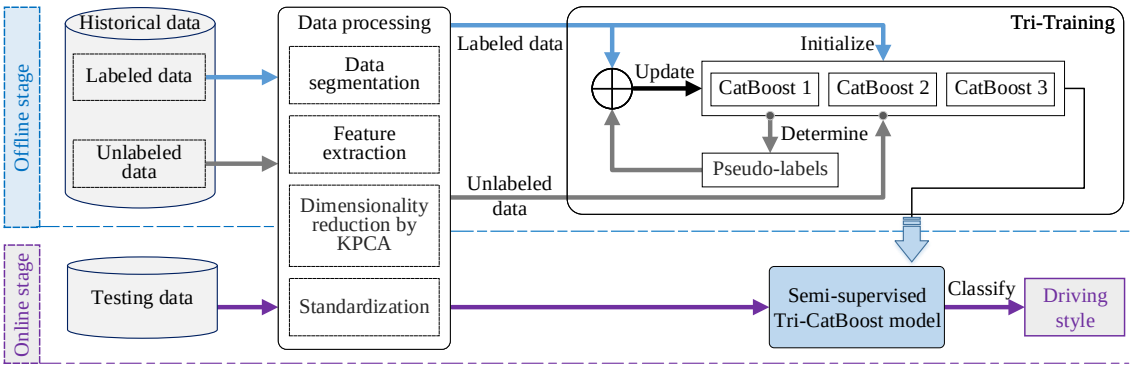

Figure 1. The framework of the proposed semi-supervised recognition method for driving style.

2.2. Feature Extraction

The velocity signal $v\,(t)$ is the basic data for driving style recognition. In addition, the acceleration $a\,(t)$ and the jerk $jerk\,(t)$ can also be taken to characterize driving styles [22]. The acceleration value reflects how quickly the speed changes and shows how the driver accelerates and decelerates. The jerk is defined as the second derivative of the velocity, which reflects the pressure intensity applied by the driver on the acceleration or brake pedal [23]. The formula is shown in Equation (1). Therefore, these signals are combined to form the raw driving signal vector $(v, a, jerk)$, which reflects the characteristics of driving styles from different aspects.

$$jerk\,(t) = \frac{d^2 v\,(t)}{dt^2} \tag{1}$$

To better quantify the driving style, various derived features are extracted from each driving signal, including the mean, maximum, minimum, lower quartile (25%), upper quartile (75%), median, kurtosis, skewness, mean absolute deviation, mean standard error, and standard deviation.

In addition, the maximum product of velocity and acceleration, and the zero-crossing rate of acceleration are also calculated as new features. The above features have been proven to be able to indicate differences in different driving styles [14]. The zero-crossing rate is defined as the change rate of the acceleration signal trend. When the signs of adjacent samples are different, zero-crossing will occur. The formula is as follows:

$$zcr = \frac{1}{L}\sum_{t=1}^{L-1} s\,(t) \tag{2}$$

$$s\,(t) = \begin{cases} 1, & \mathrm{sgn}\,(a\,(t)) \neq \mathrm{sgn}\,(a\,(t-1)) \\ 0, & else \end{cases} \tag{3}$$

where $\mathrm{sgn}\,(a\,(t))$ denotes the sign of acceleratioon signal $a\,(t)$, and L is the length of $a\,(t)$.

In summary, a total of $11 \times 3 + 2 = 35$ features are constructed to characterize the driving style.

2.3. KPCA for Dimensionality Reduction

To reduce the coupling and redundancy between the extracted features, KPCA is employed to perform feature dimensionality reduction in this paper. KPCA is a typical method for non-linear dimensionality reduction. Using the kernel technique, it maps the linear inseparable input space to the linearly separable high-dimensional feature space. Then, the feature space is projected onto the low-dimensional subspace by the standard PCA. In this way, KPCA overcomes the shortcoming that the standard PCA can only realize linear dimensionality reduction. And it is well suited for dimension reduction of driving style features [24].

Assume that the data is $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, and x_i is a d-dimensional vector. The data is supposed to have been centralized. To deal with the linear inseparable data, KPCA needs to find a nolinear mapping function $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^D (D > d)$. The mapping function maps the original input space to the high-dimensional feature space, making the mapped data linearly separable. The covariance matrix of the mapped samples is calculated by

$$\mathbf{C} = \Phi(\mathbf{X}) \Phi(\mathbf{X})^{\mathrm{T}} = \sum_{i=1}^{n} \phi(x_i) \phi(x_i)^{\mathrm{T}} \tag{4}$$

The eigenvalue λ_k and the eigenvector v_k are given by

$$\mathbf{C} v_k = \lambda_k v_k \tag{5}$$

From Equations (4) and (5), we have

$$v_k = \sum_{i=1}^{n} \phi(x_i) \alpha_i^k \tag{6}$$

where $\alpha_i^k = \frac{1}{\lambda_k} \phi(x_i)^{\mathrm{T}} v_k$.

Then, by substituting Equation (6) into Equation (5), we can get

$$\sum_{i=1}^{n} \phi(x_i) \phi(x_i)^{\mathrm{T}} \sum_{j=1}^{n} \phi(x_j) \alpha_j^k = \lambda_k \sum_{i=1}^{n} \phi(x_i) \alpha_i^k \tag{7}$$

In order to avoid direct calculation of the inner product on the feature space, the kernel function is introduced:

$$\kappa(x_i, x_j) = \phi(x_i)^{\mathrm{T}} \phi(x_j) \tag{8}$$

Thus, Equation (7) can be simplified to

$$\mathbf{K} \alpha^k = \lambda_k \alpha^k \tag{9}$$

where $\mathbf{K}$ denotes the kernel matrix, and $(\mathbf{K})_{ij} = \kappa(x_i, x_j)$. Besides, $\alpha^k = \left(\alpha_1^k, \alpha_2^k, \ldots, \alpha_n^k\right)$.

Similarly to the standard PCA, KPCA performs eigenvalue decomposition on the kernel matrix $\mathbf{K}$. Then, the eigenvalues are sorted. The eigenvector corresponding to the first d' eigenvalues form the projection matrix $\mathbf{V}^* = (v_1, v_2, \ldots, v_{d'})$, which is the solution of KPCA.

For a sample x, after KPCA dimensionality reduction, its projection in the low-dimensional space is z. The coordinate of its k-th dimension is as follows:

$$z^k = \sum_{i=1}^{n} \alpha_i^k \kappa(x_i, x) \tag{10}$$

2.4. CatBoost Classifier

In recent years, the variants of gradient boosting algorithms have developed rapidly. Under the framework of gradient-boosting decision tree (GBDT), three major implements have emerged: eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost. XGBoost achieves massive parallelism and has wide industrial applications [25]. LightGBM effectively improves the calculation efficiency [26]. However, both XGBoost and LightGBM have an inherent problem of prediction shift. By solving this problem, CatBoost has become a more promising algorithm than XGBoost and LightGBM in terms of accuracy and generalization ability. Specifically, CatBoost is an ensemble of symmetric decision trees, whose symmetry structure endows it fewer parameters, faster training and testing, and higher accuracy. In addition, CatBoost replaces the gradient estimation

method of the traditional gradient boosting algorithm with ordered boosting, thereby reducing the bias of the gradient estimation and improving the generalization capability [21]. These advantages motivate us to choose CatBoost as the base classifier of the proposed driving style recognition method.

Given a dataset $D = \{(x_k, y_k)\}_{k=1}^{n}$, where $x_k = \left(x_k^1, \ldots, x_k^{d'}\right)$ is a d'-dimensional feature vector and $y_k \in \mathbb{R}$ is the corresponding label. The symmetric decision trees are constructed by recursively partitioning the entire feature space. Assume that the feature space $\mathbb{R}^{d'}$ of CatBoost is divided into J disjoint regions (tree nodes). Each region (leaf of the tree) has a corresponding value b_j, which is the estimated value of the predicted class label. A decision tree h can be written as a superposition of estimated values of all regions:

$$h(x) = \sum_{j=1}^{J} b_j \mathbb{1}_{\{x \in R_j\}} \tag{11}$$

where, function $\mathbb{1}_{\{x \in R_j\}}$ is a indicator function:

$$\mathbb{1}_{\{x \in R_j\}} = \begin{cases} 1 & if \quad x \in R_j \\ 0 & otherwise \end{cases} \tag{12}$$

In the gradient boosting process, a series of approximate functions $F^t : \mathbb{R}^{d'} \to \mathbb{R}$ are established to minimize the expected loss $\mathcal{L}(F) := \mathbb{E}L(y, F(x))$ in a greedy manner:

$$F^t = F^{t-1} + \alpha h^t \tag{13}$$

where α is the step size and h^t is a tree which is selected from a series of H functions to minimize $\mathcal{L}(F)$ in the t-th iteration:

$$\begin{aligned} h^t &= \text{argmin}_{h \in H} \mathcal{L}\left(F^{t-1} + h\right) \\ &= \text{argmin}_{h \in H} \mathbb{E}L\left(y, F^{t-1}(x) + h(x)\right) \end{aligned} \tag{14}$$

The least-squares function is used as a loss function for the most part. And the negative gradient step is used to solve the minimization problem. Consequently, Equation (14) is transformed as follows:

$$h^t = \text{argmin}_{h \in H} \mathbb{E}\left(-g^t(x, y) - h(x)\right)^2 \tag{15}$$

where $g^t(x, y) := \frac{\partial L(y,s)}{\partial s}\Big|_{s=F^{t-1}(x)}$. After N iterations, we get a series of approximate functions $F^t(t = 0, 1, \ldots)$ and sum them to get the final model:

$$F(x) = \sum_{t=1}^{N} h^t \tag{16}$$

It should be mentioned that the goal of the first tree h^1 is y, while the latter trees h^t regard the residuals r^t of the targets y and the estimated results $h^{t-1}(x)$ of the previous model as their goals.

In the standard gradient boosting procedure, the expectations in Equation (15) are unknown and are usually approximate as below using the same dataset D:

$$h^t = \text{argmin}_{h \in H} \frac{1}{n} \sum_{k=1}^{n} \left(-g^t(x_k, y_k) - h(x_k)\right)^2 \tag{17}$$

Therefore, the classification model F^t at each boosting step relies on the target values of all training samples which are also used to build the previous model F^{t-1}. It indicates the target leakage, resulting in the prediction shift of the learned model. To put it in another way, the target y_k has been used in the previous steps of boosting. Thus, the conditional distribution $F^{t-1}(x_k)\,|x_k$ on the training sample x_k is shifted from the conditional distribution $F^{t-1}(x)\,|x$ on the testing sample x. The prediction shift will affect the generalization ability of the model and the performance of driving style classification. It is worth noting that all existing gradient boosting methods are confronted with such issues.

Based on the principle of the ordered boosting, CatBoost improves the standard gradient boosting process and implements no bias boosting. The principle of the ordered boosting is as follows: suppose x_k is ordered according to a random permutation σ. The model M_i is trained using the first i samples of σ, and a total of m models $M_1, \ldots, M_n$ are obtained. At each step, the residual of the current sample is obtained through subtracting the target value of the i-th sample and $M_{i-1}(x_i)$. In this way, no target y_k is revealed in the previous steps of boosting, and CatBoost can achieve a high generalization performance.

2.5. Tri-CatBoost for Semi-Supervised Driving Style Recognition

In this paper, Tri-CatBoost is proposed to classify driving styles as aggressive, normal and drowsy. On the one hand, Tri-Training is a semi-supervised learning method based on disagreement. It can use the information contained in a large amount of unlabeled data to improve the classification performance of driving style with little labeled data [20]. On the other hand, CatBoost is a classifier with excellent classification performance and generalization ability. Combining their advantages, Tri-CatBoost is able to reduce the requirements for labels and improve the ability to classify driving styles. The proposed Tri-CatBoost is trained according to the semi-supervised learning strategy Tri-Training. The mechanism is as follows: Three CatBoost base classifiers are initialized with labeled data firstly; then, the pseudo-labels of unlabeled data are generated by these base classifiers; the pseudo-labeled data and the raw labeled data are used to iteratively update the base classifier until they no longer change. Finally, three base classifiers are fused into a strong classifier according to a simple voting method.

Let L denote the original labeled sample set and U denotes the unlabeled sample set. First, three labeled training sets are obtained by conducting Bootstrap Sampling on L. Three CatBoost classifiers, say CB_1, CB_2 and CB_3, are initially generated from these three training sets. After that, the "minority obeying majority" strategy is used to generate the pseudo-labels of the samples. Specifically, assuming that the prediction results of classifiers CB_2 and CB_3 on an unlabeled sample x are consistent, the sample can be labeled. Then, the pseudo-labeled sample is added to the training set of another classifier CB_1 so that it can be updated. After continuous iterations, all three classifiers no longer change. It should be noted that in each round, the pseudo-labeled samples in the previous round will be retreated as unlabeled. Finally, an integrated classifier is obtained through the simple voting method implemented on three classifiers.

However, if the classification of CB_2&CB_3 on x is wrong, a noisy mislabeled sample will be obtained. If mislabeled samples are added to the training set, the performance of CB_1 will be affected. Therefore, the newly labeled samples should satisfy the following condition:

$$\left|L \cup L_1^t\right| \left(1 - 2\frac{\eta_L |L| + \tilde{e}_1^t \left|L_1^t\right|}{\left|L \cup L_1^t\right|}\right)^2 > \left|L \cup L_1^{t-1}\right| \left(1 - 2\frac{\eta_L |L| + \tilde{e}_1^{t-1} \left|L_1^{t-1}\right|}{\left|L \cup L_1^{t-1}\right|}\right)^2 \tag{18}$$

where L_1^t and L_1^{t-1} are the sets of newly labeled samples of CB_1 in the t-th round and the $(t-1)$-th round, respectively. $L \cup L_1^t$ and $L \cup L_1^{t-1}$ are the training sets of CB_1 in the t-th round and the $(t-1)$-th round, respectively. $\tilde{e}_1^t$ denotes the upper limit of the classification error of CB_2&CB_3 in the t-th round, which is estimated using the original labeled sample set U. Equation (18) is equivalent to

$$0 < \frac{\tilde{e}_1^t}{\tilde{e}_1^{t-1}} < \frac{\left|L_1^{t-1}\right|}{\left|L_1^t\right|} < 1 \tag{19}$$

Only if Equation (19) is satisfied, can L_1^t be taken as the newly-added training set of CB_1. Otherwise, the random sampling is performed on L_1^t, and its size after sampling is

$$s = \left\lceil \frac{\tilde{e}_1^{t-1} \left| L_1^{t-1} \right|}{\tilde{e}_1^t} - 1 \right\rceil \tag{20}$$

If L_1^{t-1} meets the condition in Equation (21), then L_1^t will satisfy Equation (19) after sampling. At this time, the samples in L_1^t can be added to the training set to update CB_1.

$$\left| L_1^{t-1} \right| > \frac{\tilde{e}_1^t}{\tilde{e}_1^{t-1} - \tilde{e}_1^t} \tag{21}$$

The pseudo-code of Tri-CatBoost algorithm for driving style recognition is shown in Algorithm 1.

Algorithm 1 The procedure of the Tri-CatBoost method for driving style recognition

Input:

L: Labeled sample set

U: Unlabeled sample set

Train three CatBoost classifiers $CB_i\,(i \in \{1,2,3\})$ by a set generated through Bootstrap Sampling. And initialize $\tilde{e}_i^0 \leftarrow 0.5$, $L_i^0 \leftarrow 0.5$, $t \leftarrow 0$
repeat
  for each CB_i **do**
    Let $L_i \leftarrow \emptyset$, $update_i \leftarrow FALSE$
    Calculate the classification error rate $\tilde{e}_i^t$ of CB_j&$CB_k\,(j,k \neq i)$
    for every $x \in U$ **do**
      if $CB_j = CB_k\,(j,k \neq i)$ **then**
        $L_i^t \leftarrow L_i^t \cup \{(x, CB_j(x))\}$
      end if
    end for
    if $0 < \frac{\tilde{e}_i^t}{\tilde{e}_i^{t-1}} < \frac{\left|L_i^{t-1}\right|}{\left|L_i^t\right|} < 1$ **then**
      $update_i \leftarrow TRUE$
    else
      if $\left|L_i^{t-1}\right| > \frac{\tilde{e}_i^t}{\tilde{e}_i^{t-1} - \tilde{e}_i^t}$ **then**
        Perform random sampling on L_i^t with the subsampling size $\left\lceil \frac{\tilde{e}_i^{t-1}\left|L_i^{t-1}\right|}{\tilde{e}_i^t} - 1 \right\rceil$
        $update_i \leftarrow TRUE$
      end if
    end if
    if $update_i = TRUE$ **then**
      retrain CB_i using $L \cup L_i^t$
    end if
  end for
  $t \leftarrow t + 1$
until none of CB_i changes

Output: An integrated classifier through the voting method on $CB_i\,(i \in \{1,2,3\})$

3. Data Analysis and Feature Processing

In this section, the driving dataset and data segmentation are described. After an in-depth analysis of driving data, feature extraction, dimensionality reduction and standardization are carried out.

3.1. UAH-DriveSet Description

The realistic public data set UAH-DriveSet is utilized as the experimental data set for driving style recognition. The data set provides a wealth of driving data collected from six different drivers

and different types of vehicles. Drivers repeated natural driving on predetermined roads. Each driver emulates three different driving styles (aggressive, normal and drowsy) to drive [27]. In total, more than 500 min of natural driving data were generated. The collected data includes GPS data, inertial sensor data and video data. For the sake of simplicity, only highway velocity data collected at the frequency of 1 Hz is used for experiments in this paper. Moreover, the driving data at the start stage of the vehicle are not taken into consideration. A total of 16,265 samples were obtained for driving style estimation.

3.2. Data Segmentation

Driving data consists of multiple trips with different styles. The driving style is consistent during each trip, including multiple continuous velocity data. According to the data analysis, the number of trips is limited, but the travel time of a single trip is long, ranging from 478 to 1069 s. Therefore, each trip is segmented with a period of 60 s, and the remaining portions that less than 30 s are discarded. The schema of the data segmentation is shown in Figure 2. After segmentation, 273 segments of data are finally obtained.

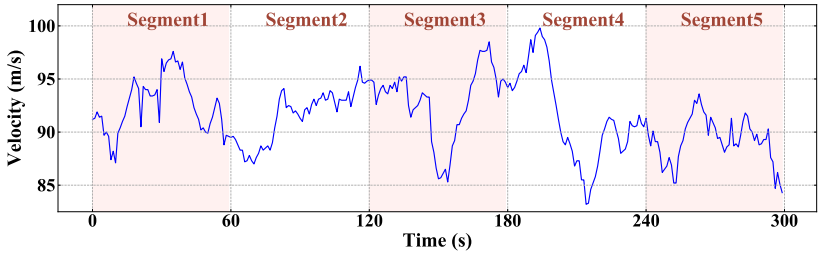

Figure 2. The schematic diagram of data segmentation.

3.3. Feature Extraction and Dimensionality Reduction

To characterize driving styles, the acceleration and jerk are calculated from the velocity signal to form the raw driving signal group together. The typical driving signals and their density distributions for different driving styles are shown in Figure 3. In the aggressive driving style, vehicles travel at high speed and are often accompanied by emergency braking and sharp acceleration. It is reflected in the high values and large fluctuations of the velocity, acceleration, and jerk signals. In addition, the density distributions of these signals are very scattered. In the drowsy driving style, the vehicle velocity is relatively low and stable, and the distributions of acceleration and jerk are concentrated and close to zero. The driving performance of the normal driving style is somewhere in between. It can be seen that velocity, acceleration and jerk can effectively reflect the differences between driving styles.

Based on raw driving signals, a total of 35-dimensional statistical features, e.g., maximum, minimum, kurtosis, skewness, and zero-crossing rate, are extracted, as described in Section 2.2. Taking the velocity signal as an example, Table 1 compares the feature values of different driving styles. It can be seen that the mean, maximum, minimum, lower quartile (25%), upper quartile (75%), median, mean absolute deviation, mean standard error, and standard deviation of the velocity signal have an obvious positive correlation with the aggressiveness of driving style. The larger the value of these features, the closer to the aggressive driving style. Moreover, the skewness of the normal driving style is negative, indicating that its probability density distribution is asymmetric and left-skewed. The skewness of the aggressive and drowsy driving styles is close to zero, indicating that the probability density distribution curve is symmetrical with respect to the average. Obviously, these features show obvious differences under different driving styles, which indicates that the extracted features can fully demonstrate driving style characteristics.

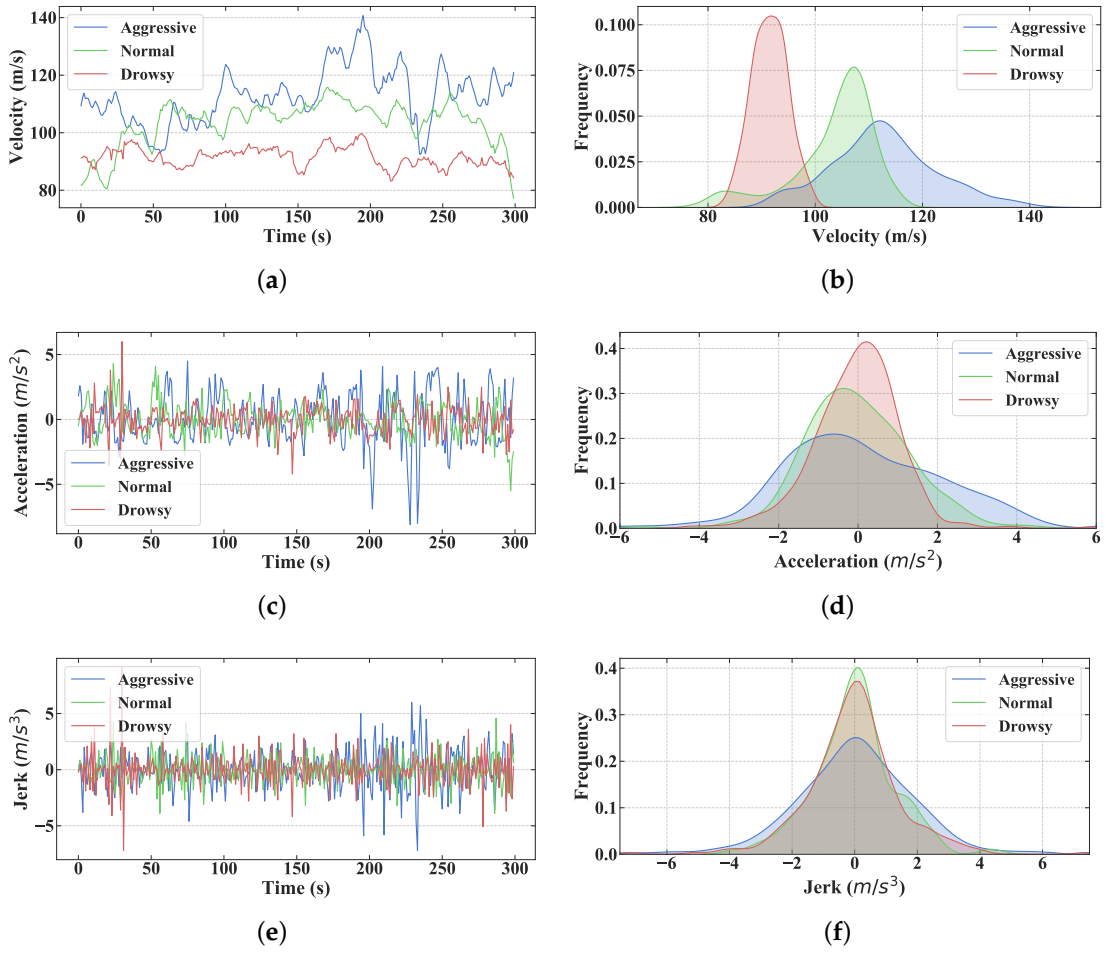

Figure 3. Typical raw driving signals of aggressive, normal and drowsy driving styles: (**a**) velocity. (**b**) distribution of velocity. (**c**) acceleration. (**d**) distribution of acceleration. (**e**) jerk. (**f**) distribution of jerk.

Table 1. Statistical feature values of velocity for different driving styles.

Feature	Aggressive Style	Normal Style	Drowsy Style
Mean	112.78	103.47	91.53
Maximum	140.70	115.80	99.80
Minimum	92.50	77.10	83.20
Lower quartile (25%)	106.80	100.30	89.10
Upper quartile (75%)	118.25	108.60	93.90
Median	112.10	105.40	91.45
Kurtosis	0.04	1.47	−0.39
Skewness	0.24	−1.34	0.03
Mean absolute deviation	7.43	5.81	2.70
Mean standard error	0.56	0.45	0.19
Standard deviation	9.64	7.78	3.31

After feature extraction, KPCA is used to reduce the dimensionality of the extracted features to eliminate the feature coupling and reduce the computational burden. In this paper, the polynomial kernel function is adopted by KPCA. In the step of dimension reduction, only labeled data can be used to determine the number of reduced features. Therefore, the number of features under different labeling ratios may be different. To set the parameter, the dimension is tuned from 1 to 34 in the condition of full labels and the best results are achieved when dimension is 16. Thus, the number of reduced features in all experiments is set to 16, which is the optimal size under full labels.

In order to improve the convergence speed and calculation accuracy of the driving style recognition model, Z-score standardization is performed on the dimension reduced features. After standardization, the data mean is 0 and the standard deviation is 1:

$$x_i^{j\prime} = \frac{x_i^j - \bar{x}^j}{\sigma^j} \tag{22}$$

where x_i^j and $x_i^{j\prime}$ denote the j-th feature value of the i-th sample before and after standardization, respectively. $\bar{x}^j$ and σ^j denote the mean and standard deviation of the j-th feature values of all samples, respectively [28].

4. Driving Style Recognition with Semi-Supervised Tri-CatBoost

In this section, the evaluation metrics of driving style recognition are first introduced. Then, a large number of experiments are carried out on UAH-DriveSet to verify the superiority of the proposed Tri-CatBoost method.

The experimental hardware environment is a PC equipped with Intel Core i5 CPU and 8 GB RAM. The software environment is Anaconda3 software and the Python version is 3.7. This paper uses the machine learning library "Scikit-learn" and the CatBoost open-source algorithm library. The data obtained by previous processing is divided at a ratio of 7:3 to obtain the training set and the test set. In order to reduce the influence of random factors, the reported results are the average of ten independent repeated trials.

4.1. Evaluation Metrics

To evaluate the performance of the proposed model, macro precision, macro recall, and AUC are used as evaluation metrics [29,30]. Let N_{ij} $(i, j \in \{1, 2, 3\})$ denote the number of samples for which the real style is the i-th style but the estimated style is the j-th style. The macro precision $macro-P$ and the macro recall $macro-R$ reflect the overall classification accuracy and the coverage, respectively:

$$macro-P = \frac{1}{3}\sum_{k=1}^{3}\frac{N_{kk}}{\sum\limits_{i=1}^{3} N_{ik}} \tag{23}$$

$$macro-R = \frac{1}{3}\sum_{k=1}^{3}\frac{N_{kk}}{\sum\limits_{j=1}^{3} N_{kj}} \tag{24}$$

The receiver operating characteristic (ROC) curves are also used to study the generalization performance of the model. Its quantitative indicator is the area under the ROC curve (AUC). The larger the AUC, the better the model performs.

4.2. Analysis of the Proposed Semi-Supervised Tri-CatBoost Method

In order to verify the superiority of the proposed driving style recognition method, the recognition results of the proposed semi-supervised learning method Tri-CatBoost, supervised learning method gradient boosting decision tree (GBDT), and unsupervised clustering method K-Means are compared. The cluster number of K-Means is three.

The ROC curves of supervised GBDT and the proposed semi-supervised Tri-CatBoost in different rates of labeled data are shown in Figure 4. The diagonal indicates the ROC curve of the stochastic guess model, and the (0,1) point corresponds to the ideal model. Due to the limited number of test samples, the ROC curves are not smooth. The ROC curves of both methods are above the diagonal and are getting closer to the (0,1) point as the rate of labeled data increases. In most cases, the ROC curve of

Tri-CatBoost can wrap that of GBDT. In addition, the AUC of Tri-CatBoost is greater than that of GBDT. This shows that the driving style classification performance of Tri-CatBoost is better than that of GBDT.

The classification results of semi-supervised Tri-CatBoost, supervised GBDT, and unsupervised K-Means in different rates of labeled data are shown in Figure 5. The percentage of labeled data was increased from 10 to 90 for testing. It is obvious that the rate of labeled data has no effect on K-Means, and the macro precision and the macro recall of K-Means are at a low level and remain unchanged. It is worth mentioning that when the percentage of labeled data is less than 20%, the macro precision and the macro recall of K-means are the highest. At this time, the labels are extremely limited. Thus, it is difficult for the supervised learning and semi-supervised learning methods to build an accurate classification model. For Tri-CatBoost and GBDT, both the macro accuracy and the macro recall show an obvious upward trend with the increase of the ratio of labeled data. Moreover, the semi-supervised Tri-CatBoost has the highest macro precision and macro recall scores overall.

In general, compared with the supervised learning and unsupervised learning, the proposed semi-supervised learning method Tri-CatBoost exhibits the best driving style recognition performance. The conclusion is quantified from AUC, macro precision and macro recall. It is mainly because that Tri-CatBoost combines the Tri-Training strategy that is able to make full use of the information in the unlabeled data and the CatBoost classifier that has a superior classification performance.

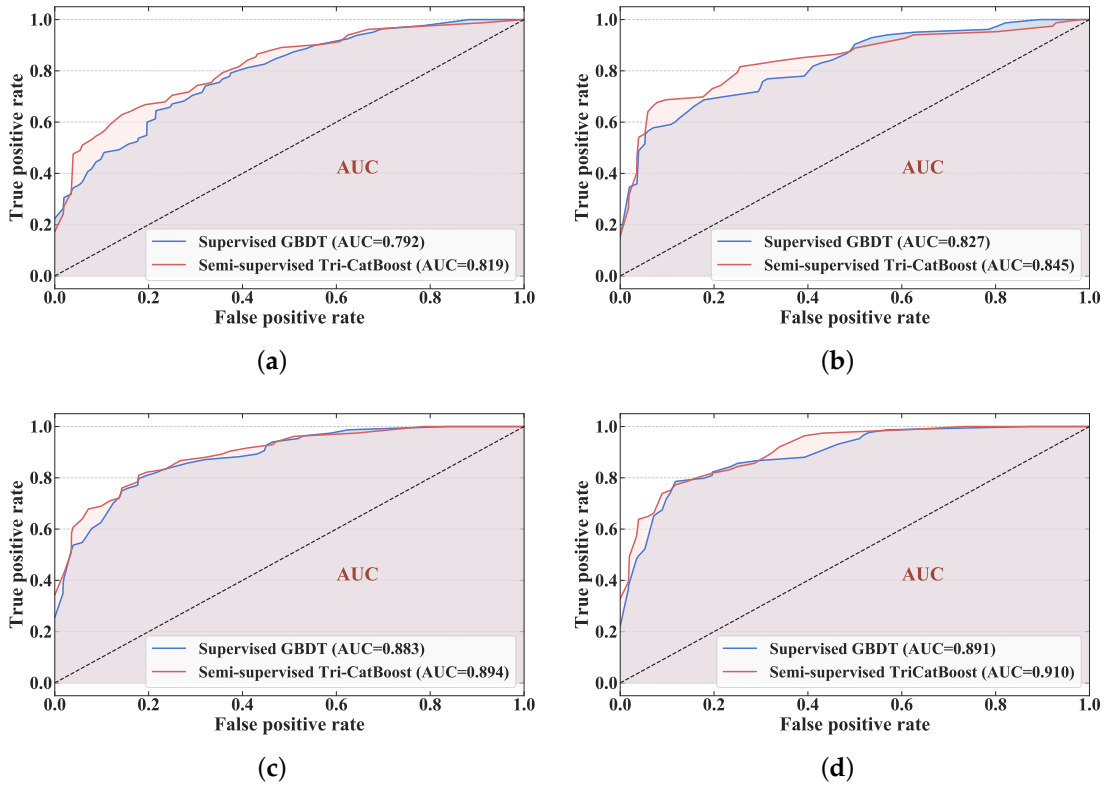

Figure 4. ROC curves of the proposed semi-supervised Tri-CatBoost and supervised GBDT: (**a**) 20% labeled data. (**b**) 40% labeled data. (**c**) 60% labeled data. (**d**) 80% labeled data.

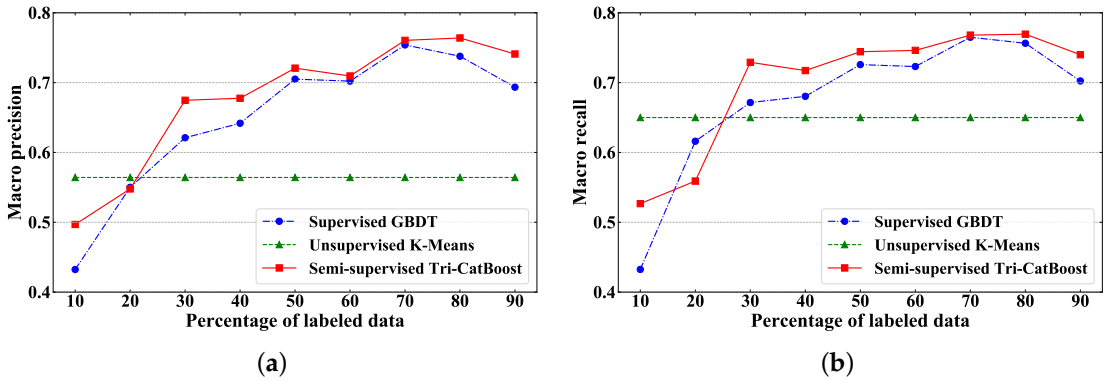

Figure 5. Classification results of semi-supervised Tri-CatBoost, supervised GBDT and unsupervised K-Means: (**a**) macro precision. (**b**) macro recall.

4.3. Analysis of the Semi-Supervised Tri-Training Strategy

To verify the ability to utilize unlabeled data to improve the classification effect and reduce label dependency, the semi-supervised strategy Tri-Training employed is compared with other two classical strategies. These two strategies are Self-Training and Co-Training. Self-Training uses a classifier to generate pseudo-labels for unlabeled samples, and then iteratively updates the classifier [31]. Co-Training uses two feature subsets to train two base classifiers respectively, which are then updated in the manner of Self-Training [32]. In order to control the variables, all the above three methods use CatBoost as the base classifier.

Figure 6 shows the ROC curves of Self-CatBoost, Co-CatBoost, and the proposed Tri-CatBoost. Comparing the ROC curves under 20% and 80% labeled data, it is found that the AUC values of all three strategies are significantly improved. The reason is that with the increase of the labeled data, all three strategies can get more definite labels, which reduces the difficulty of the base classifier to give correct pseudo-labels. When the label ratio is 20%, the ROC curve of the proposed Tri-CatBoost can almost completely wrap the ROC curves of Self-CatBoost and Co-CatBoost. In the case of other label ratios (40%, 60%, 80%), the ROC curves of the three strategies intersect. By comparing the AUC values, it is concluded that the overall classification ability of the proposed Tri-CatBoost exceeds those of Self-CatBoost and Co-CatBoost. The result fully illustrates the effectiveness of Tri-Training.

Figure 7 compares the macro precision and macro recall of these strategies at different ratios of labeled data. It can be seen from the figure that the changing trends of the macro precision and the macro recall are not completely consistent. The macro precision improves with the increase of the labeling ratio and tends to be stable when the labeling ratio is greater than 70%, while the macro recall is stable when the labeling ratio is greater than 30%. Overall, the Tri-CatBoost model trained using the Tri-Training strategy performs best in both macro precision and macro recall.

In summary, compared with other semi-supervised learning strategies Self-Training and Co-Training, Tri-Training shows the best classification performance. It is attributed to the fact that the Tri-Training strategy can make full use of the differences between three classifiers to mine unlabeled data information more effectively.

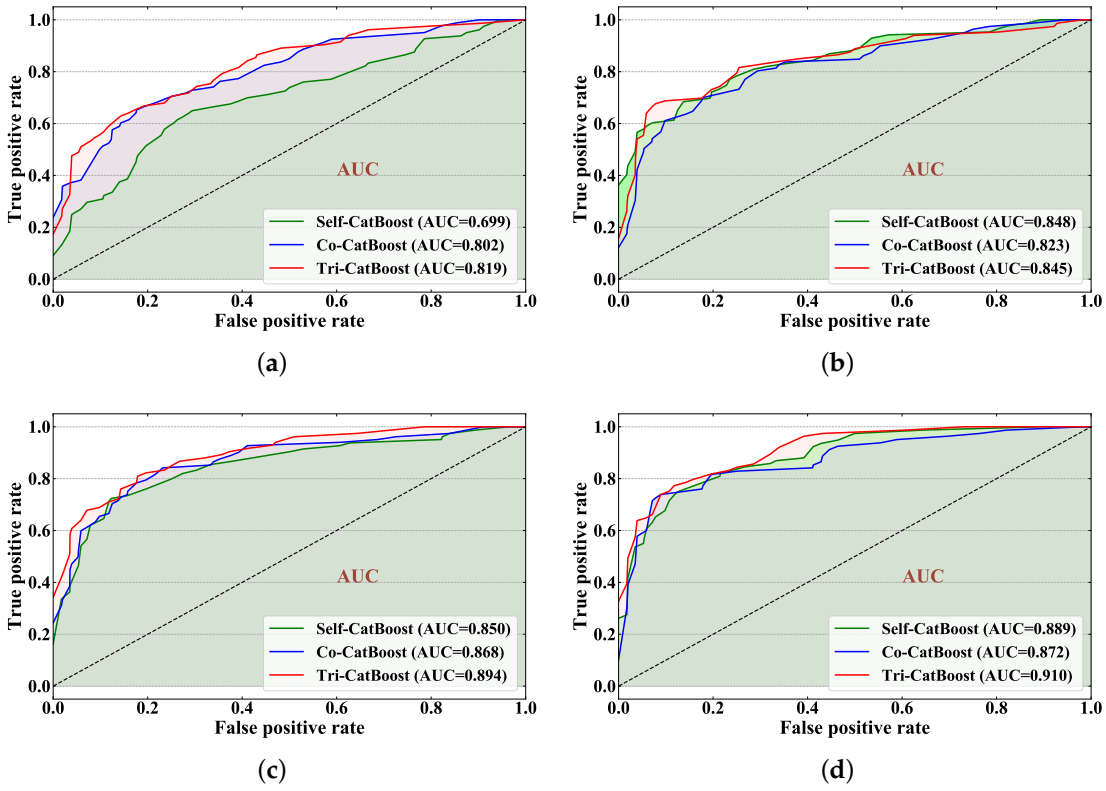

Figure 6. ROC curves of different semi-supervised learning strategies Tri-Training, Self-Training and Co-Training, using CatBoost as the base classifier: (**a**) 20% labeled data. (**b**) 40% labeled data. (**c**) 60% labeled data. (**d**) 80% labeled data.

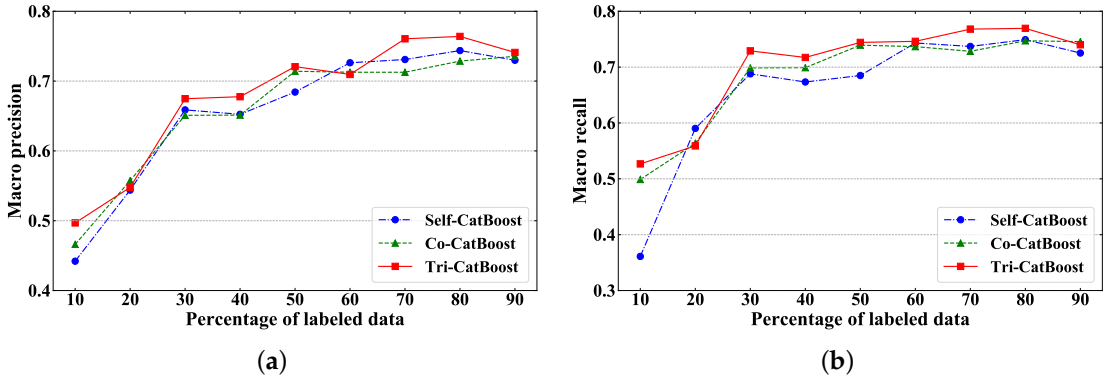

Figure 7. Classification results of different semi-supervised learning strategies Tri-Training, Self-Training and Co-Training, using CatBoost as the base classifier: (**a**) macro precision. (**b**) macro recall.

4.4. Analysis of Base Classifier CatBoost

In order to verify the classification ability for driving style, the base learner CatBoost is compared with other classifiers, including random forest (RF), decision tree (DT), multilayer perceptron (MLP), and GBDT. All of the above classifiers are trained using the Tri-Training strategy, and the classification results are shown in Figure 8. It can be seen that CatBoost can achieve higher macro precision and macro recall than other base classifiers, especially when the rate of labeled data is higher than 30%. This is mainly due to its excellent learning ability and generalization ability.

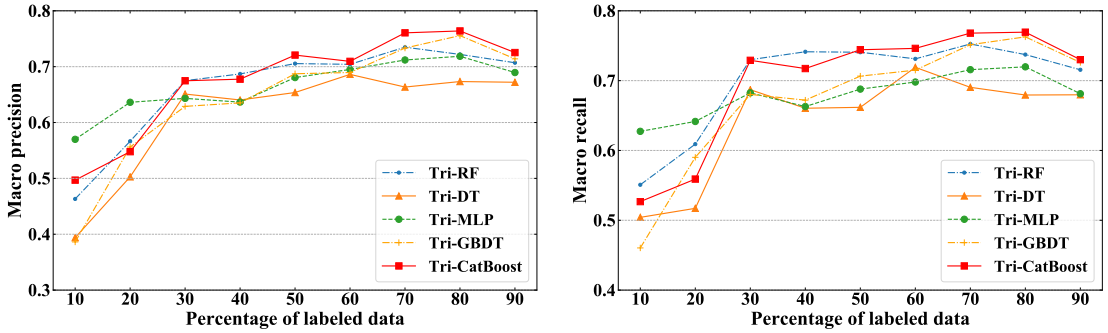

Figure 8. Classification results of CatBoost and other base classifiers, using the Tri-Training strategy: (**a**) macro precision. (**b**) macro recall.

4.5. Analysis of Classification Results for Different Driving Styles

Taking the case where the rate of labeled data is 50% as an example, the classification results of the proposed semi-supervised Tri-CatBoost method for different styles are analyzed. Figure 9a shows the ROC curves of the Tri-CatBoost classification result for each driving style. The ROC curve of the aggressive style is very close to the point of (0,1). The ROC curve of the normal style is closest to the diagonal. In other words, the classification performance for the normal style is the worst. The ROC curve of the drowsy style is between the ROC curves of the aggressive style and the normal style. The AUC of the aggressive style is the highest, the drowsy style is the second, and the normal style is the lowest.

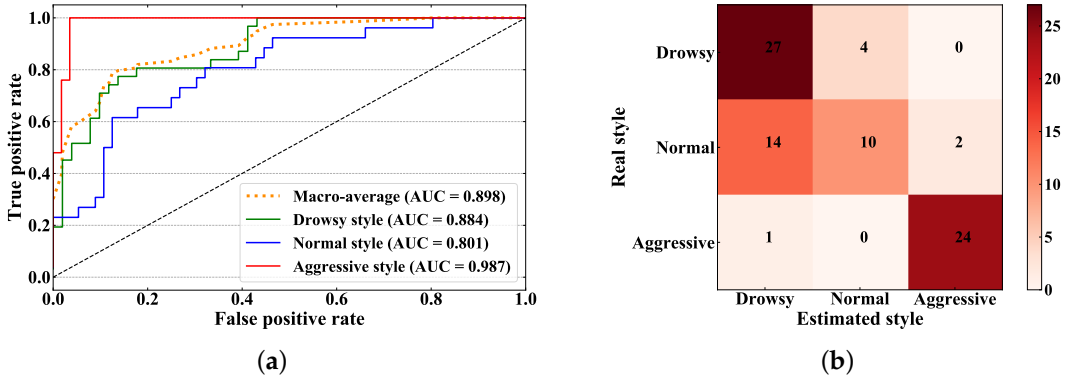

Figure 9. Classification results of the proposed Tri-CatBoost with 50% labeled data: (**a**) ROC curve. (**b**) confusion matrix.

Figure 9b shows the confusion matrix of classification results of the proposed Tri-CatBoost method. The total number of testing trips is 82, and the overall false positive rate is 25.6%. 76.2% of false positives are caused by misclassifying the normal styles, most of which are classified to the drowsy style. This is mainly because some segments in the normal style are similar to that of the drowsy style, which brings great difficulties to classification. It is of great significance to reduce the false positive rate of these segments. The misclassification related to the aggressive style is the lowest, which is mainly because that the extracted features can effectively reflect the distinctive characteristics of the aggressive style.

5. Discussion

In order to verify the superiority of the proposed Tri-CatBoost method, various experiments were designed and the results were analyzed in detail. In Section 4.2, the classification results of unsupervised K-means, supervised GBDT and semi-supervised Tri-CatBoost are compared. Taking the

case where the labeling ratio is 50% as an example, the macro precision of these methods is 0.564, 0.705, 0.721, respectively. The macro recall is 0.650, 0.726, 0.744, respectively. It is concluded that the proposed method Tri-CatBoost has better classification performance under limited labels than unsupervised learning and supervised learning. In Section 4.3, the classification results of Tri-Training and other semi-supervised learning strategies, Self-Training and Co-Training, are compared. The macro precision of these methods is 0.684, 0.714, 0.721, respectively when labeling ratio is 50%. The macro recall is 0.685, 0.739, 0.744, respectively. It is concluded that the Tri-Training strategy can mine unlabeled data information more effectively than other semi-supervised strategies. In Section 4.4, CatBoost and other base classifiers are compared, e.g., RF, DT, MLP, GBDT. The macro precision of CatBoost is 1.5% higher than that of RF, 6.7% higher than that of DT, 4.0% higher than that of MLP, and 3.4% higher than that of GBDT trained by Tri-Training. The macro recall of CatBoost is also higher than other methods. It is concluded that CatBoost has excellent classification and generalization ability, which is very suitable as a basic classifier for driving style recognition. In Section 4.5, the classification results of different driving styles are compared. The overall false positive rate is 25.6%. The classification result of the aggressive style is the best and that of the normal style is the worst.

6. Conclusions

In this paper, a new data-driven semi-supervised driving style recognition method Tri-CatBoost is proposed. The proposed method combines the semi-supervised learning strategy Tri-Training and the ensemble learning classification algorithm CatBoost. It can make full use of the data resources to reduce the dependence on data labels and improve recognition accuracy effectively. Specifically, a large number of statistical features are extracted from the raw driving signals to characterize the driving style. KPCA dimension reduction and Z-score standardization are performed on features to reduce coupling and accelerate model convergence, respectively. Then, the CatBoost algorithm with strong learning ability and generalization ability is used to establish base classifiers according to the processed labeled data. Moreover, the Tri-Training strategy is used to update the base classifiers using unlabeled data to further improve classification performance. Finally, the three base classifiers are fused into the final driving style recognition model. A series of experiments are performed on a public data set and the effectiveness of the proposed method is proved. Compared with other supervised and unsupervised learning methods, the proposed semi-supervised method exhibits the best classification performance. It has broad industrial application prospects. In the future, we will further study the driving style recognition under various working conditions.

Author Contributions: Formal analysis, X.Z.; Methodology, K.D.; Project administration, W.L.; Resources, F.J.; Software, Y.C.; Validation, J.P.; Visualization, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant Nos. 61672539, 61672537, 61873353).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Martinez, C.M.; Heucke, M.; Wang, F.Y.; Gao, B.; Cao, D. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Trans. Intell. Transp.* **2017**, *19*, 666–676. [CrossRef]
2. Lv, C.; Hu, X.; Sangiovanni-Vincentelli, A.; Li, Y.; Martinez, C.M.; Cao, D. Driving-style-based codesign optimization of an automated electric vehicle: A cyber-physical system approach. *IEEE Trans. Ind. Electron.* **2018**, *66*, 2965–2975. [CrossRef]
3. Suzdaleva, E.; Nagy, I. Two-layer pointer model of driving style depending on the driving environment. *Transp. Res. B Meth.* **2019**, *128*, 254–270. [CrossRef]
4. Jachimczyk, B.; Dziak, D.; Czapla, J.; Damps, P.; Kulesza, W. IoT on-board system for driving style assessment. *Sensors* **2018**, *18*, 1233. [CrossRef] [PubMed]

5. Lv, C.; Xing, Y.; Lu, C.; Liu, Y.; Guo, H.; Gao, H.; Cao, D. Hybrid-learning-based classification and quantitative inference of driver braking intensity of an electrified vehicle. *IEEE Trans. Veh. Technol.* **2018**, *67*, 5718–5729. [CrossRef]
6. Rajan, B.; McGordon, A.; Jennings, P. An investigation on the effect of driver style and driving events on energy demand of a PHEV. *World Electr. Veh. J.* **2012**, *5*, 173–181. [CrossRef]
7. Ding, N.; Ma, H.; Zhao, C.; Ma, Y.; Ge, H. Data anomaly detection for internet of vehicles based on traffic cellular automata and driving style. *Sensors* **2019**, *19*, 4926. [CrossRef] [PubMed]
8. Wang, W.; Xi, J.; Chen, H. Modeling and recognizing driver behavior based on driving data: A survey. *Math. Probl. Eng.* **2014**, *2014*, 245641. [CrossRef]
9. Xue, Q.; Wang, K.; Lu, J.J.; Liu, Y. Rapid Driving Style Recognition in Car-Following Using Machine Learning and Vehicle Trajectory Data. *J. Adv. Transp.* **2019**, *2019*, 9085238 . [CrossRef]
10. Huang, B.; Cohen, K.; Zhao, Q. Active anomaly detection in heterogeneous processes. *IEEE Trans. Inform. Theory* **2018**, *65*, 2284–2301. [CrossRef]
11. Huang, B.; Cohen, K.; Zhao, Q. Active anomaly detection in heterogeneous processes. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; Volume 65, pp. 3924–3928.
12. Wang, W.; Xi, J. A rapid pattern-recognition method for driving styles using clustering-based support vector machines. In Proceedings of the American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016; pp. 5270–5275.
13. Brombacher, P.; Masino, J.; Frey, M.; Gauterin, F. Driving event detection and driving style classification using artificial neural networks. In Proceedings of the IEEE International Conference on Industrial Technology (ICIT), Toronto, ON, Canada, 22–25 March 2017; pp. 997–1002.
14. Xie, J.; Zhu, M. Maneuver-Based Driving Behavior Classification Based on Random Forest. *IEEE Sens. Lett.* **2019**, *3*, 1–4. [CrossRef]
15. Sun, B.; Deng, W.; Wu, J.; Li, Y.; Zhu, B.; Wu, L. Research on the Classification and Identification of Driver's Driving Style. In Proceedings of the 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; Volume 1, pp. 28–32.
16. Zhou, Z.H.; Li, M. Semi-supervised learning by disagreement. *Knowl. Inf. Syst.* **2010**, *24*, 415–439. [CrossRef]
17. Guo, J.; Liu, Y.; Zhang, L.; Wang, Y. Driving behaviour style study with a hybrid deep learning framework based on GPS data. *Sustainability* **2018**, *10*, 2351. [CrossRef]
18. Ozgul, O.F.; Cakir, M.U.; Tan, M.; Amasyali, M.F.; Hayvaci, H.T. A Fully Unsupervised Framework for Scoring Driving Style. In Proceedings of the International Conference on Intelligent Systems (IS), Madeira, Portugal, 25–27 September 2018; pp. 228–234.
19. Feng, Y.; Pickering, S.; Chappell, E.; Iravani, P.; Brace, C. Driving style analysis by classifying real-world data with support vector clustering. In Proceedings of the 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), Singapore, 3–5 September 2018; pp. 264–268.
20. Zhou, Z.H.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Environ.* **2005**, *11*, 1529–1541. [CrossRef]
21. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 6638–6648.
22. Vaitkus, V.; Lengvenis, P.; Žylius, G. Driving Style Classification Using Long-Term Accelerometer Information. In Proceedings of the 19th International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, Poland, 2–5 September 2014; pp. 641–644.
23. Murphey, Y.L.; Milton, R.; Kiliaris, L. Driver's style classification using jerk analysis. In Proceedings of the IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, Nashville, TN, USA, 30 March–2 April 2009; pp. 23–28.
24. Wang, Q. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv* **2012**, arXiv:1207.3538.
25. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

26. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
27. Romera, E.; Bergasa, L.M.; Arroyo, R. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 387–392.
28. Deng, K.; Zhang, X.; Cheng, Y.; Zheng, Z.; Jiang, F.; Liu, W.; Peng, J. A Remaining Useful Life Prediction Method with Automatic Feature Extraction for Aircraft Engines. In Proceedings of the 18th IEEE International Conference on Trust, Security And Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019; pp. 686–692.
29. Van Asch, V. *Macro- and Micro-Averaged Evaluation Measures*; Tech. Report; CLiPS: Belgium, Brussel, 2013.
30. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
31. Dong, C.; Schäfer, U. Ensemble-style self-training on citation classification. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–3 November 2011; pp. 623–631.
32. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.