Weight Feedback-Based Harmonic MDG-Ensemble Model for Prediction of Traffic Accident Severity

Byung-Kook Koo ¹, Ji-Won Baek ² and Kyung-Yong Chung ¹,*

¹ Division of AI Computer Science and Engineering, Kyonggi University, Suwon 16227, Korea; byuguki@kyonggi.ac.kr
² Department of Computer Science, Kyonggi University, Suwon 16227, Korea; jwbaek@kyonggi.ac.kr
* Correspondence: dragonhci@gmail.com

Abstract: Traffic accidents are emerging as a serious social problem in modern society but if the severity of an accident is quickly grasped, countermeasures can be organized efficiently. To solve this problem, the method proposed in this paper derives the MDG (Mean Decrease Gini) coefficient between variables to assess the severity of traffic accidents. Single models are designed to use coefficient, independent variables to determine and predict accident severity. The generated single models are fused using a weighted-voting-based bagging method ensemble to consider various characteristics and avoid overfitting. The variables used for predicting accidents are classified as dependent or independent and the variables that affect the severity of traffic accidents are predicted using the characteristics of causal relationships. Independent variables are classified as categorical and numerical variables. For this reason, a problem arises when the variation among dependent variables is imbalanced. Therefore, a harmonic average is applied to the weights to maintain the variables’ balance and determine the average rate of change. Through this, it is possible to establish objective criteria for determining the severity of traffic accidents, thereby improving reliability.

Keywords: ensemble; weight-feedback; traffic accident; accident severity; harmonic; mean decrease Gini coefficient

1. Introduction

Most traffic accidents are caused by vehicles, and thus severe injuries and deaths outnumber slight-moderate injuries. According to the World Health Organization (WHO), globally about 1,350,000 people die due to traffic accidents every year, and the annual number of deaths is on the rise [1]. The influences on traffic accidents include human factors, such as a driver’s physical defects, driving habits, and safety consciousness material factors including poor vehicle maintenance and deterioration, environmental factors, including weather, time, and traffic enforcement; and physical factors, including the road type and condition [2,3]. Therefore, it is necessary to deal with traffic accidents by considering their causes and influencing factors. Indicators of traffic accidents’ severity, such as the number of vehicles involved and the number of deaths and injuries, can be found at an accident site. The severity of a traffic accident is decided after expert analysis based on diverse factors. If a traffic accident’s severity is assessed immediately, the scale of support from emergency vehicles such as ambulances and others can be arranged efficiently. Therefore, a method for deriving traffic accident severity objectively and immediately must be developed.

With the number of car drivers increasing, there have been more demands for analyzing the actual extent of individual traffic accidents and the related potential risk factors. In addition, the development of IoT (Internet of Things) leads to a massive amount of data being generated and collected in real-time. Accordingly, big data processing technologies have been developed to find significant information among road traffic data. A typical technology for big data processing is machine learning, which is classified as supervised or unsupervised learning. Machine learning models include a support vector machine,
decision tree, and k-nearest neighbor’s algorithm. In these models, new knowledge and information are obtained from pre-processed data through classification, regression, correlation analysis, and other methods. In a single machine learning model, it is difficult to reflect a variety of data features, such as data distributions and patterns. Consequently, overfitting occurs because the training data is a subset of the actual data in some cases [4,5]. Therefore, we proposed a traffic accident prediction model using [6] Gaussian-kernel support vector machines (SVMs), such as used by B. Sharma et al. This solves the problem of data properties reflecting the characteristics of the data used through a Gaussian kernel SVM. However, this has problems that are limited to the data used in the study. In addition, Baek et al. [7] proposed a ContextDNN (Context Deep Neural Network) model using multiple regression for risk prediction. The proposed method divides the data into a certain training, testing, and validation ratio to solve the over-fitting problem. Because this divides limited data, it is difficult to use when the amount of data is insufficient. It also predicts risk by discovering causal relationships through regression analysis. However, if data are of a categorical type, it is difficult to find the rate of change of independent variables to dependent variables. Therefore, it is necessary to conduct an efficient analysis with a small amount of data and to develop a method for finding the causal relationships of categorical data according to the rate of change of independent variables to dependent variables. Hsu et al. [8] proposed a method for analyzing the severity of traffic accidents at intersections using a logistic regression model. The proposed method uses nine independent variables, including the month of the accident, the period of the accident, the weather condition and condition, the road type, the road surface condition, and the traffic control type. Accordingly, using a logistic regression model, it was analyzed that accident time, road surface condition, accident type, and vehicle type had a significant influence on the severity of traffic accidents. However, the severity method using the logistic regression model is used when the category is in binary format. Traffic accidents need to be predicted from multi-categorical data because analyzing only in binary format creates ambiguity. Therefore, this study proposes a weighted feedback-based harmonic MDG (Mean Decrease in Gini)-ensemble model for the prediction of traffic accident severity by using traffic accident data. It considers objective criteria for determining the severity of traffic accidents and the characteristics of various data and solves the over-fitting problem to reflect traffic accident severity. For performance evaluation, the proposed ensemble algorithm is compared with Random-Forest, KNN (K-Nearest Neighbor Algorithm), and XGbTree (eXtreme Gradient Boosting Tree) as single algorithms in terms of classification accuracy. In addition, an accuracy comparison based on the application of the mean decrease in Gini, coefficient harmonic-mean conformity, AUC of proposed model and comparison with conventional models are analyzed.

This paper is comprised of the following: Section 2 describes the variable selection using the MDG coefficient and the classification methods that use ensembling. Section 3 describes the weight feedback-based harmonic MDG-ensemble model for the prediction of traffic accident severity; Section 4 describes the performance evaluation; and Section 5 presents the conclusion of the study.

2. Related Work

2.1. Variable Selection Using Mean Decrease Gini Coefficient

Causal interrelationships occur frequently in everyday life. For example, changes in weather, speeding, and a driver’s carelessness, and other factors in the traffic area lead to traffic accidents. To analyze the causal interrelationships of these data, independent and dependent variables were established. In the causal analysis, variables of low importance for dependent variables can generate errors that reduce the quality of the analysis. Consequently, variables with low importance must be eliminated to improve the quality of the causal analysis. The variable importance analysis method is used as the process of finding interrelationships between dependent and independent variables. Variable importance analysis is aimed at finding independent variables’ contribution to dependent variables [9].
Therefore, by analyzing the importance of independent variables, it is possible to remove unnecessary independent variables. A method exists to use the MDG coefficient in the variable importance analysis. The MDG coefficient is the scale used to determine the level of decrease in the classification impurity of a feature in the decision tree-based random forest model. This is the mean of impurity values of the features for each tree in random forest. The larger the MDG coefficient, the more it influences the classification of dependent variables [10,11]. Accordingly, classification is conducted based on the features for a particular node. As a result of the classification, each parent node generates its child node. Therefore, the impurity of the parent node and child node can be calculated. A decrease in impurity represents the decrease in a child node’s impurity relative to its parent node. In a single node-based tree structure, the lower impurity is the better classification for features.

Han et al. [12] proposed a variable selection method using the random forest-based MDG coefficient and the Mean Decrease in Accuracy. The proposed method designs a random forest model for ten microarray data sets, in which the Mean Decrease in Accuracy and MDG coefficient are derived. Through this, the Mean Decrease in Accuracy and the MDG coefficient are derived. Ranking is applied to each of the variables based on the Mean Decrease in Accuracy and the MDG coefficient, and thus the top 50% of variables only are selected. Through the iteration of the process, the variable found when Error Rate and CPU (Central Processing Unit) Time are the lowest is finally selected. This shows more accuracy and requires less time than the conventional variable selection method. Given that, if the MDG coefficient is used, it is possible to select the variables for prediction effectively.

2.2. Classification Methods Using Ensemble

Classification is capable of predicting the category of dependent variables according to the specific criteria of data [13,14]. A prediction model for classification has little difference depending on data features. There are many different learning methods related to classification. It is important to select a learning method that fits data features. A single learning method has difficulty reflecting the complicated features of data. To solve this problem, an ensemble technique is applied. An ensemble technique uses the given data to merge multiple classifiers and construct a model [15]. It can be divided into Bagging (Bootstrap AGGregATING) [16], Boosting [17], and Stacking [18] according to the merging methods of single models. Bagging utilizes the result of Voting, which is the process of extracting samples multiple times, training each model, and totaling a result. In addition, because it adjusts the value extracted from each sample with the use of a median value, it solves the over-fitting problem. Voting is classified into hard-voting, soft-voting, and weighted-voting according to the use of weighting and method of a particular model. Hard-voting is the method of totaling the prediction values of models and selecting the value with the most votes [19]. Soft-voting is the method of adding up probabilities of models and selecting the class with the highest value [20]. Weighted-voting is the method of assigning a different weight to each of the single models in the ensemble. Accordingly, in the ensemble model, none of the models have the same contribution [21]. The random forest learning model is an ensemble model to which the decision tree-based hard-voting Bagging method is applied. Figure 1 shows the bagging structure of the parallel classifier connection using voting.

As shown in Figure 1, Bootstrap, the method of recovering and randomly extracting the samples with the same size from the given data, is applied to generate training data. In this manner, each model is trained. The final result is obtained by the voting of the trained model’s values. Based on the result, classifiers are connected in parallel for integration [16]. Boosting is a technique for generating a weak classifier the performance of which is similar to that of the classifier drawn by random guessing [17]. Figure 2 shows the boosting technique of using adaptive weights.
As shown in Figure 1, Bootstrap, the method of recovering and randomly extracting the samples with the same size from the given data, is applied to generate training data. In this manner, each model is trained. The final result is obtained by the voting of the trained model’s values. Based on the result, classifiers are connected in parallel for integration [16]. Boosting is a technique for generating a weak classifier the performance of which is similar to that of the classifier drawn by random guessing [17]. Figure 2 shows the boosting technique of using adaptive weights.

As presented in Figure 2, during boosting the classifiers run sequentially. In this case, the input data are extracted by Bootstrap, as they are with Bagging. Boosting repeatedly gives a high weight to a wrong answer, and a low weight to a correct answer. Because Boosting improves the error of a previous classifier, it has better performance than Bagging. However, it is easily over-fitted on the input data. In addition, Stacking generates a classification model with the best performance through the merging of multiple classifiers, such as SVM, random forest, and KNN [18]. Figure 3 shows the stacking method using the performance of the classification prediction result.

As shown in Figure 3, the Stacking method splits the entire data corpus into test data and training data. The classification result for the test data is drawn from the multiple base models that had learned the training data. Therefore, the training data are reconstructed and the final classifier learns to draw the classification results for the test data. Through test data, a classifier with good performance was used. For this reason, the advantages of each classifier are applied, and their weaknesses are mitigated. Accordingly, studies have been conducted on an appropriate number of single models to improve the performance of ensemble models in diverse fields, including face recognition and remote sensing [15,22]. Xiao et al. [23] proposed an SVM and KNN ensemble model for the detection of traffic accidents. The proposed model uses a single SVM and KNN for traffic accident data. The ensemble model is designed using the Bagging method. This gives the weight of the
individual single detection models. This results in the model with the highest weight being selected. Compared to a single detection model, this improves the accuracy of traffic accident detection. However, since it does not consider the rate of change of independent variables for dependent variables, it is difficult to recognize accident severity. Therefore, to consider severity, dependent variables must be predicted according to the rate of change in independent variables.

3. Weight Feedback Based Harmonic MDG-Ensemble Model for Predicting Traffic Accident Severity

To predict traffic accident severity, the weight feedback-based harmonic MDG-ensemble model undertakes three steps. The first is to pre-process the traffic data. The data are based on traffic accident information provided by the Korean Road Traffic Authority [24]. Through pre-processing, missing values and outliers are eliminated. Attributes with duplicate meanings are integrated. Based on the variables of data on very serious traffic accidents, the severity of accidents is generated as a dependent variable, and the text and numerical variables are categorized. The second step is to select variables using the mean decrease in Gini (MDG). For the extraction of the MDG coefficient, a decision tree-based learning model capable of calculating the impurity of variables is used. By using pre-processed data, the random forest classifier is modeled for traffic accident risk classification. From this, the MDG of independent variables for traffic accident severity is extracted. In the third step, traffic accident severity is predicted through a weight feedback-based harmonic-MDG ensemble. As for the weights used for modeling, the harmonic-mean-based change rate is taken into account and is used as feedback. Accordingly, a harmonic-MDG ensemble model with excellent performance was assembled. Figure 4 shows the weight feedback-based harmonic MDG-ensemble model process for the prediction of traffic accident severity.
there are various types of traffic accidents, such as vehicle-to-vehicle, vehicle-to-person, and vehicle-to-bicycle accidents. In addition, the degree of injury, number of deaths, and costs of accident handling are different. Accordingly, the level of traffic accident severity is subjectively predicted based on the severity of a driver’s injuries, the number of deaths, or the extent of vehicle damage. The severity attributed by a subjective opinion lacks supporting evidence and has low reliability. Thus, objective criteria must be set for determining severity according to the types of traffic accidents. To eliminate unnecessary data from the analysis and to increase efficiency, pre-processing is applied. The traffic accident information data used in this study is based on the 2012 to 2018 serious traffic accident information provided by the Korean Road Traffic Authority [24]. This includes twenty-seven variables (e.g., date and day of occurrence, number of deaths, and number of injuries) in 30,913 transactions. However, no variables related to traffic accident severity were found. Therefore, following the accident severity calculation method used by the Korean Road Traffic Authority [24], traffic accident severity variables are generated and categorized. This is calculated by multiplying the number of deaths and serious injuries by 70%, and the number of casualties by 30%. Accordingly, if the value of a traffic accident’s severity is below 2, the severity level is categorized as ‘2’; if between 2 and 3, it is categorized as ‘2’; if more than 3, it is categorized as ‘3’. The data consists of numerical and text data and must be converted into numerical and categorical data through pre-processing. Data with duplicate meanings or the data that are unconvertible to numerical data like ‘occurrence place type (city or county)’ are eliminated. Table 1 shows the pre-processing results of traffic accident information data, consisting of a variable name, data type, category, and meaning.

As shown in Table 1, 13 variables and 18,700 transactions were extracted through pre-processing. Under ‘Traffic Accident Severity’, ‘1’ means ‘Slight-moderate’; ‘2’ means ‘Serious’; ‘3’ means ‘very serious’. ‘Year’ is categorized according to the year of occurrence. ‘Violation_law’ a large category of law violations is categorized between ‘0’ and ‘2’: ‘0’ means a pedestrian’s negligence; ‘1’ means a driver’s violation; ‘2’ means poor maintenance of the vehicle. ‘Road_type’ a large category of road type is categorized between ‘0’ and ‘4’: ‘0’ means crossroad; ‘1’ means others/undefined; ‘2’ means one way only; ‘3’ means parking lot; ‘4’ means railroad crossing. However, based on the pre-processing results, it is difficult to identify the factors that influence severity. For this reason, the important variables that influence severity are extracted. Therefore, to prevent over-fitting of data analysis, it is divided into 70% of training data, 20% of test data, and 10% of validation data.
Table 1. Pre-processing results of Traffic accident information data.

| No | Variable Name       | Data Type  | Category | Description            |
|----|---------------------|------------|----------|------------------------|
| 1  | Traffic Accident   | Categorical| 1        | Slight-moderate        |
|    | Severity            |            | 2        | Serious                |
|    |                     |            | 3        | very serious           |
| 2  | Year                | Categorical| 2012     | Year of accident       |
|    |                     |            | 2013     | occurrence             |
|    |                     |            | 2014     |                         |
|    |                     |            | 2015     |                         |
|    |                     |            | 2016     |                         |
|    |                     |            | 2017     |                         |
|    |                     |            | 2018     |                         |
| 12 | Violation_law      | Categorical| 0        | Pedestrian’s negligence|
|    |                     |            | 1        | Driver’s violation of law|
|    |                     |            | 2        | Poor maintenance of vehicle|
| 13 | Road_type          | Categorical| 0        | Crossroad              |
|    |                     |            | 1        | Others/Undefined       |
|    |                     |            | 2        | One way only           |
|    |                     |            | 3        | Parking lot            |
|    |                     |            | 4        | Railroad crossing      |

3.2. Variable Selection According to Importance Using Mean Decrease in Gini Coefficient

Traffic accidents are influenced by diverse factors, such as speed, poor driving, and road conditions. Traffic data include the variables unnecessary for severity prediction or weakly influential variables. Therefore, accurately to predict traffic accident severity, influential and highly important variables are selected. Since pre-processed data include categorical data, it is difficult to select variables based on regression analysis. For this reason, a modeling method according to data types must be developed [25,26]. In this study, the importance of variables is derived by applying the decision tree-based random forest. Variable importance refers to selecting independent variables that are important for predicting changes in independent variables [27,28]. For variable selection, the MDG coefficient which is the mean of the decrease in impurity values for specific variables in a decision tree-based model [29] was used. Traffic accident severity was set as a dependent variable to extract its MDG coefficient and the other 12 variables were set as independent variables [30,31]. The MDG coefficient for each of the independent variables was derived by a classifier. The relative importance of the dependent variable was determined in comparison to the MDG coefficient. Figure 5 shows the variable importance using the MDG coefficient. The horizontal axis represents the mean decrease in Gini. The vertical axis represents the independent variables of traffic accident information data.

However, since Korea has four distinct seasons, it is necessary to consider the variables for year and month. In addition, traffic congestion occurs according to the time of commute to work, and the accident rate is high. Therefore, it is necessary to consider time. As a result of the variable importance when using the MDG coefficient in Figure 5, relatively low importance values are also extracted. If the value of the MDG coefficient is less than 500, effective analysis is not possible because the number of variables selected as MDGs decreases rapidly; therefore, a cut-off threshold of 500 was established. Table 2 shows the final result of the variable importance using the MDG coefficient.
Figure 5. Variable importance using the MDG coefficient.

Table 2. The final result of the variable importance using the MDG coefficient.

| Variable Name   | MDG      | Variable Name   | MDG      |
|-----------------|----------|-----------------|----------|
| Hour            | 1000.54  | Date            | 665.693849 |
| Casualty_type   | 1302.854 | Violation_law   | 557.599871 |
| Month           | 1058.541 | Light           | 667.584  |
| Attacker_type   | 1512.122 | Location        | 1256.67162 |

As shown in Table 2 of variable’s importance, eight variables were extracted. The MDG coefficient of Hour is 1000.54. Consequently, for traffic accident data, ‘Hour’ is judged to be the important variable in terms of severity classification. This is because accidents often occur at night or during rush hours [3,24].

3.3. Weight Feedback Based Harmonic MDG-Ensemble Model

The data for predicting traffic accident severity are numerical or categorical. Accordingly, it is important to build an appropriate model to analyze these data-types efficiently. Configurable models include machine learning-based models or deep learning-based models. In a machine learning-based model, the user directly inputs the features of the data and finds the features, and the model finds pattern learning by itself. Accordingly, new knowledge and information can be obtained. On the other hand, in the case of a deep learning-based model, the model performs self-learning without user intervention from feature extraction to pattern discovery. It also extracts results from complex and vast data. However, since it has the characteristics of a black box, there is a disadvantage in that it is possible to predict the analysis result of the data through the result without an intermediate process. In this paper, in order to improve the efficiency of classification, a model is constructed by extracting variables necessary for analysis through MDG coefficients. Therefore, a machine learning-based model is constructed. To improve the performance of multi-categorical classification, classification algorithms, such as decision tree, K-nearest neighbor (KNN), and SVM, are merged to improve the performance of multi-categorical classification. For an ensemble model, single models are merged according to specific criteria, and a new model is generated for the same dependent variable [32,33]. The proposed weight feedback based harmonic MDG-ensemble model uses the variables drawn using the MDG coefficient. Figure 6 shows the structure of the weight feedback-based harmonic MDG-ensemble model.
Figure 6. The structure of the weight feedback based harmonic MDG-ensemble model.

In Figure 6, the model has three steps. The first is to construct an ensemble model. The pre-processed traffic accident data is split into 70% of training data, 20% of test data, and 10% of verification data. Various single classification models are trained with the training data. Afterward, the ensemble model is generated using the weighted-voting ensemble method that applies these single models. The second step compares the actual and predicted values of the ensemble model to the test data. If the actual and predicted values match, this increases the “Correct” of every single model and the predicted value is the same as that of the ensemble model. Conversely, if the actual and predicted values mismatch, this increases the “Incorrect” of every single model and the predicted value matches that of the ensemble model. In the last step, the MDG values of the independent variables used in every single model, “Correct” and “Incorrect” are applied to extract the weight of every single model, and thereby weight feedback is provided. Thus, the classification performance of the ensemble model is improved. For weight feedback, a harmonic mean is applied to the MDG values. The distribution of the MDG values is biased. Therefore, the product of the MDG values has no special meaning, such as the ratio.

To solve the problem, the harmonic mean is applied to the MDG coefficient. The arithmetic mean is used when variables have common criteria [34]. In traffic accident data, however, factors such as time, month, position, year, and type of accident have no common criteria. Therefore, this division is not applied to calculate the mean. The geometric mean is used when it is possible to analyze the ratio of common change in variables [35]. Because traffic accident data include categorical data, it is impossible to determine the rate of change in variables. On the contrary, for the harmonic mean, variable importance is not biased but remains balanced, and thus the average ratio of change [36,37] can be found. Therefore, the harmonic mean is applied to the MDG values for each of the variables, and traffic accident severity is predicted. In the case of the harmonic mean, the inverse number of each factor is calculated through the arithmetic mean and is then converted into the inverse number again, and the resulting value is calculated. The value of the harmonic MDG coefficient increases indefinitely and thus the scope is reduced using normalization. This provides weighted feedback for an ensemble model. The weight for reducing the prediction error is readjusted. Equation (1) shows the weight feedback in the harmonic MDG ensemble model. \( FW \) is the feedback weight. \( N \) is the number of independent variables used for the generation of an ensemble model. \( i \) is used to distinguish an individual single model used in the ensemble model. \( j \) is used to count the independent variables used in the single model, and MDG represents the MDG value for each variable. \( \alpha \) and \( \beta \) show the amount of increase and decrease in weight of a single model, respectively, calculated using the weighted feedback algorithm.
Inacc respectively are used to determine whether the corresponding single model has the highest contribution to the accurate or inaccurate prediction of an ensemble model. 

\[
H_{MDG_i} = \left( \frac{1}{N} \sum_{j=1}^{N} \frac{1}{MDG_{ij}} \right)^{-1}
\]

\[
FW_i = \text{Weight}_i + H_{MDG_i}(\alpha \times \text{Acc}_i - \beta \times \text{Inacc}_i)
\]

(1)

As shown in Equation (1), for weight feedback, the sum of the harmonic mean of the MDG values, and the contribution of every single model is added to the previous weight. Accordingly, through weight feedback, the error is reduced, and the performance of the harmonic MDG-ensemble model is improved. A weight can be updated by considering the over-fitting problem arising when an ensemble model is created and the importance of independent variables used for the generation of a single model. Algorithm 1 shows the weight feedback-based harmonic MDG-ensemble algorithm. Input data is the pre-processed traffic accident data, and the output data is the weight updated by feedback.

**Algorithm 1. Weight feedback based harmonic MDG-ensemble algorithm.**

**Input:** Pre-processed Traffic Accident Data → TAD\[j\]

**Output:** Updated Weight using Feedback → FW

`Num_Model ← Number of single models for ensemble`

`Correct [Num_Model] ← 0`

`Incorrect [Num_Model] ← 0`

`Accurate [Num_Model] ← 0`

`Inaccurate [Num_Model] ← 0`

`Weight [Num_Model] ← 1`

Split TAD into into 70% of training data, 20% of test data, and 10% of verification data

for `k` is the number of split TAD do

Trains data of TAD sets is number of `k`, and the rest is a dataset for testing or validation.

Trains Model [Num_Model] with single classification method using train data

Makes Ensemble model WFEM using Model, Weight

for `j` = 1 to length of train fold do

Pred_e ← Prediction of WFEM for j-th transaction of test data

Real ← Real Label of j-th transaction of test fold

for `i` = 1 to Num_Model do

Pred_s ← Prediction of Model[i] for j-th transaction of test data

if (Real == Pred_e)///Under accurate prediction

if (Pred_e == Pred_s)///Prediction of Model[i] is same as WFEM

Correct[i]++

else if (Real ≠ Pred_e)///Under inaccurate prediction

if (Pred_e ≠ Pred_s)

Incorrect[i]++

endfor

endfor

Idx_Correct ← index of maximum of correct

Idx_Incorrect ← index of maximum of incorrect

Accurate [Idx_Correct] ← 1

Inaccurate [Idx_Incorrect] ← 1

FW_i = Weight_i + H_{MDG_i}(\alpha \times \text{Acc}_i - \beta \times \text{Inacc}_i)

Accurate ← 0

Inaccurate ← 0

endfor

return

The weight feedback-based harmonic MDG-ensemble algorithm process of Algorithm 1 is based on the generation of an ensemble model and weight feedback. First, the variable for saving the correct or incorrect value is initialized to '0'. Also, the variable that determines the contribution of accurate or inaccurate predictions of the ensemble model is initialized...
to ‘0’. The initial weight of each model is set to ‘1’. Next, the pre-processed traffic accident data is split into 70% of training data, 20% of test data, and 10% of verification data set. This prevents over-fitting for model learning. The weights were adjusted according to the prediction results of the ensemble models and single models for the test fold. For every single model, if the prediction values of the ensemble and single models are consistent with each other while predicting accurately, the single model’s Correct increases; if they are consistent under inaccurate prediction, the single model’s Incorrect increases. The Correct and Incorrect tallies of all single models are compared, and thus the Accurate and Inaccurate of the model with the largest value for either is set to ‘1’. The difference in values multiplied by the weight increase (Alpha) and reduction (Beta) multiplied by the Accurate and Inaccurate, respectively, and the calculated value is multiplied by the harmonic-mean of the MDG. After that, the calculated value is added to an existing weight and the result is offered as feedback. Finally, Accurate and Inaccurate are set to ‘0’ for the weight feedback to be made for the rest of the fold. These tasks are repeated as often as the number of folds split TAD. Thus, the error of the ensemble model is improved.

4. Result and Performance Evaluation

4.1. Traffic Accident Severity Prediction System

Traffic accidents arise due to diverse causes, such as weather, time of day, driver’s condition, and driving career. The severity of traffic accidents is predicted based on vehicle damage and the number of deaths. Traffic accident severity is subjectively determined by an expert, its reliability is low, and it is difficult to make an immediate decision based thereon. To solve this problem, this study applies the weight feedback-based harmonic MDG-ensemble model for the prediction of traffic accident severity to establish a traffic accident severity prediction system. The operating system and hardware specifications of the system are as follows: Window10 Education, Intel(R) Core(TM) i5-4690 3.50 GHz, 32 GB RAM. Figure 7 shows the traffic accident severity prediction system using the weight feedback-based harmonic MDG-ensemble model. When a traffic accident occurs, a user collects traffic accident factors. Without pre-processing the factors, the user fills in each of the items in the system, similar to a survey. In the model information, the user selects a classification model version. When the user completes data input, the system inputs the data into the traffic accident severity prediction model. A categorical value of traffic accident severity is then derived.

Figure 7. Traffic accident severity prediction system using weight-feedback based harmonic MDG-ensemble model.
The traffic accident severity prediction system in Figure 7 consists of traffic accident factor collection, position preview, model selection, traffic accident severity derivation, and model information. The traffic accident factors selected through the MDG are the categories or values attributable to the eleven variables. A user can find information on an input position visually through the system. In addition, the traffic accident severity prediction system updates a prediction model when data collection is completed or a data update occurs. Therefore, a user can select and use models with different versions. Based on the collected information, the model derives the severity of a traffic accident. If the value is 0, it means slight-moderate injuries; if it is 1, it means serious injuries; if it is 2, it means very serious injury. In addition, the system provides information on the single models that constitute the traffic accident severity prediction model.

For example, as shown in Figure 7, traffic accident severity data are collected according to regions. The MDG extracts the importance of each variable that meets the 500 threshold. Subsequently, a classification model is selected and the updated weight is calculated using the harmonic MDG-based weight feedback. As a result, the traffic accident’s severity is derived from an ensemble model with calculated weights. Thus, users can find the severity of traffic accidents in a region.

4.2. Performance Evaluation

The data used for the performance evaluation is the record of fatalities in traffic accidents from the Korea Road Traffic Authority. This is open data composed of road risk index calculated through the fusion analysis of traffic accident information, information on frequent accident areas due to traffic weakness and accident characteristics, and data such as past traffic accidents, weather, and accidents. Through traffic accident information, it is possible to extract various types of information such as where an accident occurred, severity, road risk index, etc., that can reduce traffic accidents and improve the efficacy of social cost reduction measures. For performance evaluation, the traffic accident death data are preprocessed and divided into 70% training data, 10% validation data, and 20% testing data. The proposed weight feedback-based harmonic MDG-ensemble model is used to predict traffic accident severity based on the ensemble of multiple classification algorithms. Confusion matrix-based accuracy, recall, precision, and F1-Score are used for classification evaluation metrics. Performance evaluation is conducted on four aspects.

The first is to compare the proposed ensemble model with the models based on random forest, KNN, and XgbTree as single algorithms. The second is to compare model accuracy according to the presence or absence of MDG-based variable selection. The third is to compare accuracy by applying arithmetic means, geometric means, and harmonic means to weights to evaluate the suitability for harmonic mean application in weight feedback. Finally, the performance evaluation of the proposed model is compared with a conventional accident prediction model. Table 3 shows the confusion matrix for the performance evaluation of traffic accident severity. This matrix is capable of presenting multi-categorical variables according to which models are used. Traffic accident severity is classified as Slight, Serious, and very serious. Pred represents a prediction value, and Real is an actual value.

| Real   | Pred  | Slight | Serious | Very Serious |
|--------|-------|--------|---------|--------------|
| Slight | a     | b      | c       |              |
| Serious| d     | e      | f       |              |
| very serious | g | h      | i       |              |

The multi-categorical confusion matrix in Table 3, makes it difficult to measure the accuracy, recall, precision, and F1-Score. Therefore, it is converted into a confusion matrix of the binary type (True and False), which consists of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP is the number correctly classified into a
particular category; TN is the number classified into a particular category but is different from an actual value; FP represents the number correctly classified into categories other than a particular category; FN represents the number classified into categories other than a particular category. For instance, for the category Slight, the TP is a, TN is e + f + h + i, FP is i, and FN is b + c + d + g. As such, TP, TN, FP, and FN are drawn for each category, and thus accuracy, recall, precision, and F1-Score are defined [39]. Equation (2) shows the accuracy. This means the accurately predicted ratio for the total prediction.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\] (2)

Equation (3) shows the recall. This means the accurate classification of an actual value in a particular category.

\[
\text{Recall} = \frac{TP}{TP + FN}
\] (3)

Equation (4) shows the precision. This means that a number’s classification in a particular category is consistent with an actual value.

\[
\text{Precision} = \frac{TP}{TP + FP}
\] (4)

Equation (5) presents F1-score using precision and recall.

\[
F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\] (5)

In the first performance evaluation, the proposed ensemble model is compared with the models based on random forest, KNN (K-Nearest Neighbor), and XgbTree as single algorithms. Table 4 shows the results of the performance comparison.

|                  | RandomForest | KNN (K-Nearest Neighbor) | xgbTree | Ours  |
|------------------|--------------|--------------------------|---------|-------|
| Accuracy         | 0.9655       | 0.8364                   | 0.9553  | 0.9686|
| Recall           | 0.8533       | 0.8738                   | 0.9743  | 0.9446|
| Precision        | 0.9830       | 0.4462                   | 0.8148  | 0.9089|
| F1-Score         | 0.9136       | 0.5907                   | 0.8875  | 0.9264|

As shown in Table 4, the proposed ensemble model has a better accuracy and F1-score than the single models. Random forest and XgbTree are excellent, respectively, in recall and precision. Random forest is a model for predicting traffic accident severity accurately. In addition, XgbTree is the model in which the prediction value of traffic accident severity is most consistent with its actual value. In terms of the F1-score as the harmonic mean of recall and precision, the proposed ensemble model has the highest value. Therefore, the proposed harmonic MDG-ensemble model has better general performance than the single models.

The second performance evaluation compares the performance according to the variable selection process method. Variable selection is the process of selecting independent variables that greatly influence a dependent variable. This helps to reduce the number of independent variables and thereby decreases the factors that negatively influence the analysis and reduce the modeling time. For performance evaluation, the comparison of the variable selection process and the accuracy and F1-Score were evaluated by comparing the proposed MDG, the multiple regression analysis methods, and the case without the variable selection process. Accordingly, it proves that MDG is more suitable for the selection of variables for predicting the severity of traffic accidents than other methods. This is the dependent variable for the severity of traffic accidents, and 16 variables are used as the independent variable. Accordingly, traffic accident severity prediction models
were generated using the proposed ensemble method. As a result of the variable selection process, the MDG model uses 11 independent variables and the multiple regression method uses 10 independent variables that satisfy a significant level of 0.5 or less. In addition, in the case of no variable selection process, the model uses 16 independent variables. The accuracy and F1-Score are used as the evaluation metrics for the prediction models. Figure 8 shows the result of comparing the performance of the predictive model according to the variable selection method. The vertical axis represents the accuracy and F1-Score, and the horizontal axis represents the variable selection method.

As a result of the evaluation in Figure 8, when MDG is used for variable selection, the performance is evaluated as best. The non-selected model has more independent variables than the Used-MDG model thus it uses relatively higher-dimensional data. If a model is trained with high-dimensional data, there are many cases to consider. If there are variables that less influence a model’s prediction value extraction, they negatively influence the prediction value extraction. In addition, the performance of the multiple regression method had the lowest evaluation. This is because the data used are categorical, so when the category of each data changes and the difference in results occur according to the change in significance level, the recall rate is evaluated as low, and accordingly, the F1-Score has the lowest evaluation. Therefore, in the process of selecting variables according to importance, the F1-Score and accuracy of the Used MDG model are improved compared to those of the non-selected and multiple regression models.

In the third performance evaluation, the suitability for the application of harmonic-mean to weight feedback was evaluated, the arithmetic mean, geometric mean, and harmonic mean were applied, and their accuracy was compared according to the importance of eleven variables extracted through the MDG coefficient. Figure 9 shows the results of an accuracy comparison between the arithmetic mean, geometric mean, and harmonic mean.

As shown in Figure 9, the harmonic-mean-based method is the most accurate. In the case of the arithmetic mean, since the eleven variables extracted through the MDG coefficient including time, the number of deaths, month, position, type of assailant, year, minute, type of accident, date, law violation, and day have no common criteria, there are no criteria for division. For this reason, when the arithmetic means are applied, the accuracy is low. In addition, it is impossible to find a constant rise ratio for the eleven variables. Therefore, when the geometric mean is applied, the accuracy is lowest. When the proposed harmonic mean is applied, the balance remains, and a changing mean ratio is determined. Therefore, it exhibits the best performance.
There is a difference in the degree of risk of an accident according to the threshold of the severity of a traffic accident. Therefore, the ROC (Receiver Operating Characteristic) curve is used to evaluate the proposed classification model according to the change of the threshold value. The ROC curve represents a curve through TPR (True Positive Rate) and FPR (False Positive Rate) at the thresholds of various classifications. In order to determine the superiority of the model through the ROC (Area Under the Curve), it is determined through the area of the AUC representing the area under the curve. AUC indicates from 0 to 1, and the larger the value, the better the performance. Figure 10 shows the ROC curve and AUC area of the proposed model. As a result of the evaluation, the area of AUC was evaluated as 0.871.

In the last performance evaluation, the proposed weight feedback-based harmonic MDG-ensemble model is compared with models proposed by Hashmienejad et al. [40] and Sharma et al. [6] to evaluate the quality of the proposed model. For the prediction of traffic accident severity, Hashmienejad used the improved NSGA-2 algorithm using traffic accident data provided by the Tehran Police Department. Sharma used a Gaussian-Kernel SVM with the traffic accident data provided by the Ministry of Road Transport & Highways.
India. Their traffic accident severity prediction results were compared using accuracy, recall, precision, and F1-Score. Table 5 shows the performance comparison between the proposed model and other models.

Table 5. Results of the performance comparison.

|                      | Accuracy | Recall  | Precision | F1-Score |
|----------------------|----------|---------|-----------|----------|
| Proposed             | 0.9686   | 0.9446  | 0.9089    | 0.9264   |
| Hashmienejad et al. [39] | 0.8820   | 0.8345  | 0.8856    | 0.8875   |
| Sharma et al. [6]    | 0.8606   | 0.7630  | 0.8114    | 0.7865   |

As shown in Table 5, the proposed model has better accuracy, recall, precision, and F1-Score than the models proposed by Hashmienejad and Sharma. Hashmienejad’s model analyzes traffic accident severity according to user preference. It extracts customized knowledge for road safety experts, such as traffic policemen and road and transportation engineers. It generates rules based on user preferences. In the model, the results differ depending on the experts’ data sets. Therefore, the model has a lower performance than the proposed model. Sharma’s model has no variable importance and variable selection processes. All analyzed data are categorical data. The Gaussian-Kernel SVM is a learning model influenced by the spatial distribution of data. Sharma’s model does not fit the features of the data to analyze. This is because it has lower performance than the one proposed. Thus, the proposed method can accurately predict traffic accident severity.

5. Conclusions

This study proposed a weight feedback-based harmonic MDG-ensemble model for the prediction of traffic accident severity. The proposed method consists of traffic accident data pre-processing, variable selection according to the MDG coefficient, and the construction of a weight feedback-based harmonic MDG-ensemble model. In the pre-processing stage, unique categories and redundant meanings were removed and integrated, and outliers and missing values were processed among traffic accident data of the Road Traffic Authority. In the variable selection step using the MDG coefficient, the MDG for the variable was extracted and the importance was compared to extract the variable. Accordingly, the efficiency of analysis was improved by removing unnecessary data for predicting the severity of traffic accidents. The weighted feedback-based Harmonic MDG-ensemble model was constructed through three classifiers that can be used for multi-category classification. Accordingly, various characteristics of data were considered and the over-fitting problem was solved. The weight update uses the harmonized average to consider the average rate of change. Through this, errors in the prediction of the constructed model were reduced. For performance evaluation, the proposed ensemble algorithm was compared with single classification algorithms in terms of accuracy and F1-score. Additionally, accuracy was compared according to the presence or absence of MDG-based variable selection. The suitability of the harmonic mean was evaluated and the proposed model was compared with other models in terms of performance. As a result, the proposed method demonstrated superior performance. Therefore, with the proposed method, it is possible reliably, objectively, and rapidly to determine traffic accident severity. In addition, the proposed data-based weight update method applies to the establishment of an ensemble model for use in other domains, as well as traffic data. The method proposed in this study was constructed through three single models to consider the characteristics of the data. In addition, important variables were extracted to predict the severity of accidents by selecting the importance of variables. However, a number of variables with overlapping meanings such as time and date were selected. Accordingly, the future work aims to increase the number of single models used to generate an ensemble model and, design a traffic accident severity prediction model with better performance.
Author Contributions: Conceptualization, B.-K.K. and J.-W.B.; methodology, K.-Y.C.; resources, B.-K.K., J.-W.B. and K.-Y.C.; validation, B.-K.K., J.-W.B. and K.-Y.C.; formal analysis, B.-K.K., J.-W.B. and K.-Y.C.; investigation, B.-K.K.; resources, B.-K.K.; data curation, B.-K.K. and J.-W.B.; writing—original draft preparation, B.-K.K. and J.-W.B.; writing—review and editing, J.-W.B. and K.-Y.C.; visualization, B.-K.K. and J.-W.B.; supervision, K.-Y.C.; project administration, K.-Y.C.; funding acquisition, K.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21CTAP-C157011-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

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

References
1. World Health Organization. Available online: http://who.int/ (accessed on 12 December 2020).
2. Baek, J.-W.; Chung, K. CNN-based health model using knowledge mining of influencing factors. Pers. Ubiquitous Comput. 2019, 1–11. [CrossRef]
3. Korea Road Traffic Authority Traffic Accident Analysis System. Available online: http://taas.koroad.or.kr/ (accessed on 12 December 2020).
4. Zhang, C.; Vinyals, O.; Munos, R.; Bengio, S. A study on overfitting in deep reinforcement learning. arXiv 2018, arXiv:1804.06893.
5. Wang, Z.; Srinivasan, R.S. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renew. Sustain. Energy Rev. 2017, 75, 796–808. [CrossRef]
6. Sharma, B.; Katiyar, V.K.; Kumar, K. Traffic accident prediction model using support vector machines with Gaussian kernel. In Proceedings of Fifth International Conference on Soft Computing for Problem Solving, 5th ed.; Pant, M., Deep, K., Bansal, J.C., Nagar, A., Das, K.N., Eds.; Springer: Singapore, 2016; Volume 1, pp. 1–10.
7. Baek, J.-W.; Chung, K. Context Deep Neural Network Model for Predicting Depression Risk using Multiple Regression. IEEE Access. 2020, 8, 18171–18181. [CrossRef]
8. Hsu, Y.T.; Chang, S.C.; Hsu, T.H. Analysis of Traffic Accident Severity at Intersection Using Logistic Regression Model. J. Eng. Res. Rep. 2020, 1–9.
9. Behnamian, A.; Millard, K.; Banks, S.N.; White, L.; Richardson, M.; Pasher, J. A systematic approach for variable selection with random forests: Achieving stable variable importance values. IEEE Geosci. Remote Sens. Lett. 2017, 14, 1988–1992. [CrossRef]
10. Hu, G.; Xi, T.; Mohammed, F.; Miao, H. Classification of wine quality with imbalanced data. In Proceedings of the 2016 IEEE International Conference on Industrial Technology (ICIT), Taipei, Taiwan, 14–17 March 2016.
11. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations sources and a solution. BMC Bioinf. 2007, 8, 25. [CrossRef]
12. Han, H.; Guo, X.; Yu, H. Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. In Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 26–28 August 2016.
13. Aggarwal, C.C.; Reddy, C.K. Data Clustering: Algorithms and Applications, 1st ed.; Chapman & Hall/CRC: Yorktown Heights, NY, USA, 2013; Volume 2.
14. Baek, J.-W.; Chung, K. Porthole Classification Model Using Edge Detection in Road Image. Appl. Sci. 2020, 10, 6662. [CrossRef]
15. Galicia, A.; Talavera-Llames, R.; Troncoso, A.; Koprinska, I.; Martinez-Álvarez, F. Multi-step forecasting for big data time series based on ensemble learning. Knowl. Based Syst. 2019, 163, 830–841. [CrossRef]
16. Aydoğmuş, H.Y.; Halil, I.; Erdal, A.E.; Erdal, H. Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models. J. Econ. Int. Finance. 2015, 7, 127–136.
17. Sabzevari, M.; Martínez-Muñoz, G.; Suárez, A. Vote-boosting ensembles. Pattern Recognit. 2018, 83, 119–133. [CrossRef]
18. Wang, Y.; Wang, D.; Geng, N.; Wang, Y.; Yin, Y.; Jin, Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Appl. Soft Comput. 2019, 77, 188–204. [CrossRef]
19. Randhawa, K.; Loo, C.K.; Seera, M.; Lim, C.P.; Nandi, A.K. Credit card fraud detection using AdaBoost and majority voting. IEEE Access 2018, 6, 14277–14284. [CrossRef]
20. Karlos, S.; Kostopoulou, G.; Kotsiantis, S. A Soft-Voting Ensemble Based Co-Training Scheme Using Static Selection for Binary Classification Problems. Algorithms 2020, 13, 26. [CrossRef]
21. Moustafa, S.; ElNainay, M.Y.; Makky, N.E.; Abougabal, M.S. Software bug prediction using weighted majority voting techniques. Alex. Eng. J. 2018, 57, 2763–2774. [CrossRef]
22. Htike, K.K. Efficient determination of the number of weak learners in AdaBoost. J. Exp. Theor. Artif. Intell. 2017, 29, 967–982. [CrossRef]
23. Xiao, J. SVM and KNN ensemble learning for traffic incident detection. *Phys. A Stat. Mech. Appl.* 2019, 517, 29–35. [CrossRef]
24. Korea Road Traffic Authority. Available online: http://koroad.or.kr/ (accessed on 12 December 2020).
25. Song, Y.Y.; Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* 2015, 27, 130–135. [PubMed]
26. Freeman, E.A.; Moisen, G.G.; Coulston, J.W.; Wilson, B.T. Random forests and stochastic gradient boosting for predicting tree canopy cover: Comparing tuning processes and model performance. *Can. J. For. Res.* 2016, 46, 323–339. [CrossRef]
27. Shi, L.; Westerhuis, J.A.; Rosén, J.; Landberg, R.; Brunius, C. Variable selection and validation in multivariate modelling. *Bioinformatics* 2019, 35, 972–980. [CrossRef] [PubMed]
28. Bermingham, M.L.; Pong-Wong, R.; Spiliopoulou, A.; Hayward, C.; Rudan, I.; Campbell, H.; Wright, A.F.; Wilson, J.F.; Agakov, F.; Navarro, P.; et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Sci. Rep.* 2015, 5, 1–12. [CrossRef]
29. Zhang, Z.; Yang, Z.; Ren, W.; Wen, G. Random forest-based real-time defect detection of Al alloy in robotic arc welding using optical spectrum. *J. Manuf. Process.* 2019, 42, 51–59. [CrossRef]
30. Tyralis, H.; Papacharalampous, G. Variable Selection in Time Series Forecasting Using Random Forests. *Algorithms* 2017, 10, 114. [CrossRef]
31. Mehta, S.; Shen, X.; Gou, J.; Niu, D. A new nearest centroid neighbor classifier based on k local means using harmonic mean distance. *Information* 2018, 9, 234. [CrossRef]
32. Kim, M.-J.; Kang, J.-S.; Chung, K. Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight. *IEEE Access* 2020, 8, 183983–183994. [CrossRef]
33. Powers, D.M.W. Evaluation: From precision recall and F-measure to ROC informedness markedness and correlation. *J. Mach. Learn. Technol.* 2011, 2, 37–63.