Prediction of Forest Fire Risk for Artillery Military Training using Weighted Support Vector Machine for imbalanced data

Ji Hyun Nam
Center for Army Analysis and Simulation, ROK Army HQs
Jongmin Mun
Yonsei University
Seongil Jo
Inha University
Jaeoh Kim (✉ jaeoh.k@inha.ac.kr)
Inha University

Article

Keywords:

Posted Date: June 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1736556/v1

Prediction of Forest Fire Risk for Artillery Military Training using Weighted Support Vector Machine for imbalanced data

Ji Hyun Nam[1], Jongmin Mun[2], Seongil Jo[*3], and Jaeoh Kim[†4]

[1]Center for Army Analysis and Simulation, ROKA HQs
[2]Department of Statistics and Data Science, Yonsei University
[3]Department of Statistics, Inha University
[4]Department of Data Science, Inha University

ABSTRACT

South Korea has been in a truce with North Korea since 1953, and conventional artillery firing training are still needed considering its geopolitical location. Based on the weather condition, the decision is made whether to conduct artillery training, which might cause forest fires. Wildfires caused by military training not only devastate forests but also worsen public opinion on national defense. While forest fires triggered by artillery training cause substantial damage, the number of occurrences is very low, making it difficult to construct a prediction model in a general manner. In this paper, the weighted support vector machine for imbalanced data is applied to predict the risk of forest fire due to artillery firing training. We employ an over-sampling technique based on a probability distribution for imbalanced data and applied a weighted support vector machine algorithm that enforces a misclassification cost of the minority class. This study not only considerably reduces the probability of forest fires occurring during conventional artillery fire training in the Republic of Korea Army (ROKA) but also encourages the development of practical approaches for wildfires prediction in countries with climates similar to the Korean Peninsula. Furthermore, our proposed method can contribute to the study of the classification model for various imbalanced data. Through Monte Carlo simulations, we demonstrate that our proposed method achieves significantly higher accuracy than traditional methods.

INTRODUCTION

Wildfires are one of the most devastating natural disasters that have a major impact on the ecosystem and economy. Wildfires are mostly caused by humans, such as campfires left unattended, the burning of debris, equipment use and malfunctions, inadvertently discarded cigarettes, and intentional acts of arson, among which unexpected causes are found. Over the past few years, not a few wildfires have been reported in South Korea due to the military artillery training. South and North Korea are technically still at war and have been in a truce since 1953. Considering the geopolitical location, conventional artillery firing training is still required. The artillery military training is planned in advance through sufficient deliberation and decision-making process. The process also takes into consideration the weather conditions on the day of the training. Nevertheless, each year, training sessions cause several forest fires that cause catastrophic environmental damage. Wildfires caused by military training not only devastate forests but also worsen public opinion on national defense. This deterioration of public opinion restricts the military training of the Republic of Korea Army (ROKA) to protect liberal democracy.

Methods for predicting forest fire risk have been studied in various ways. Traditional field-survey methods, such as Canadian Forest Fire Weather Index (FWI), are among the most widely used index for wildfire prediction. The index considers six components(Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, Initial Spread Index, Buildup Index, Fire Weather Index, Daily Severity Rating) that account for the effects of the fuel moisture and weather conditions on fire behavior. Several organizations [1, 2] around the world adopted the FWI system. Besides FWI, there are many other fire risk index systems such as National Fire Danger Rating System (NFDRS), Forest Fire Danger Index (FFDI), and Grassland Fire Danger Index (GFDI). The main drawback of these systems is that they use empirically derived features and equations [3, 4, 5]. In forest fire prediction, it is extremely complicated to derive a precise mathematical model explaining the complex relationship between variables, and thus several approximations are used[6]. Over the last several years, data-driven approaches have proven to be extremely useful in vast applications [7, 8, 9, 10, 11, 12] including forest fire predictions. [10] used artificial neural network and logistic regression to forecast the fire risk. [11] provided insight into the use of Random Forest (RF), Boosting Regression Trees (BRT), and Support Vector Machines (SVM) for wildfire risk assessment. Recently several studies exploited deep learning

[*] Co-corresponding author: Associate Professor, Department of Statistics, Inha University. E-mail: statjs@inha.ac.kr
[†] Corresponding author: Assistant Professor, Department of Data Science, Inha University. E-mail: jaeoh.k@inha.ac.kr

based methods for forest fire prediction [13, 14, 15].

Because of the accumulation of meteorological massive data and advanced computational capabilities to properly handle them, data-driven approaches, such as machine learning algorithms, particularly classification methods, are very beneficial in predicting wildfires. However, most machine learning algorithms are developed under the assumption that the underlying data have a similar number of observations within each class. In practice, however, there are many cases when the distribution of classes is skewed the aforementioned assumption is often no longer valid. Many binary categorical data in the real world are collected and stored as disproportionate data with significant differences in the populations of the two groups. For example, signals and images collected by military alert systems are mostly normal or animal signals, and few anomalies such as enemy infiltration are collected. In addition, in the case of diagnostic tests for certain diseases, most of the test results are negative, while positive patients are collected very rarely. In such cases, accurate prediction of minority groups is more important than misclassification of minority groups. The conventional classification methods have a problem of greatly reducing the classification accuracy of minority groups because the classifier is biased toward the majority to increase the overall classification accuracy. The data with skewed distribution of classes is called imbalanced data[16, 17].

Imbalanced data is a problem that occurs in a wide variety of fields[18, 19, 20, 21, 22, 23, 24]. Many practical cases of imbalanced data in these fields have motivated the development of various methodologies[25, 26, 27, 28, 29, 30, 31, 32] that address the class bias issues. [33] provided deep analyses of learning rate from skewed data for neural network training. [16] reviewed metrics and algorithm-level approaches for imbalanced data. [32] contributes an in-depth insight into intrinsic characteristics of data. Imbalanced data learning algorithms mostly employ one of three strategies: data augmentation-based methods, algorithm-based methods, and hybrid methods. Data augmentation methods take a data pre-processing approach and focus on balancing the data by modifying the underlying distribution of the classes. After this pre-processing, the balanced data is then fed into classical machine learning algorithms. The balancing can be achieved via under-sampling the majority classes, or over-sampling the minority classes, or using combination of these two methods. In general over-sampling methods increases the total dataset size and thus increases training time. Moreover, it can cause over-fitting to the current training dataset, and the trained model might not generalize well on new, unseen data [34]. Therefore over-sampling needs to be done in an intelligent manner. One of the widely used over-sampling method is the Synthetic Minority Oversampling Technique (SMOTE) [25]. Instead of merely replicating an existing minority class instances, the SMOTE applies linear interpolation to them and synthetically generates new minority class samples. Han et al [35] introduced a modified version of SMOTE which limits generation of additional samples to near the borderline of the class, in which the classification between the classes becomes more difficult. Bunkhumpornpat et al [27] proposed safe-region-SMOTE which synthesizes minority classes by introducing the safe region, which considers how many majority samples are nearby. Near-Miss algorithm was proposed by Mani et al [36]. The algorithm uses K-nearest neighbors and considers distance from the minority sample to determine the removal of the majority sample. Beckmann et al [37] introduced another k-nearest neighbors based method called KNN-Und that removes majority class samples based also on the count of neighbors. Barandel et al [38] used K-nearest neighbors to discard misclassified samples from the majority class. Kubat et al [29] proposed One-sided selection method using Tomek links [22]. Algorithm level methods do not modify the data distribution but instead the decision algorithm is designed such that more attention is given to the minority classes. Typically these algorithms introduce penalty or weight term to decrease the bias towards majority class. Cost-sensitive learning methods take into account the cost of misclassification. Turney et al [39] proposed ICET which uses misclassification cost to evaluate fitness function of genetic algorithm. Ling et al [40] designed decision tree algorithm with misclassification cost. Hybrid methods are combinations of data-level and algorithm-level methods applied to imbalanced data to decrease the bias towards the majority classes. Liu et al [41] designed hybrid algorithms, namely EasyEnsemble and BalanceCascade, which use combination of under-sampling and ensemble learning to tackle imbalanced data classification task. Ramentol et al [31], Gao et al [42], Alberto et al [43] proposed several hybrid methods based on SMOTE algorithm. SMOTEBoost introduced by Chawla et al [44] uses combinations sampling techniques and ensemble methods. Mease et al [45] developed JOUS-Boost algorithm using AdaBoost with over/under-sampling methods.

The purpose of this study is to develop a prediction model for the forest fires that often occur during the artillery fire training of in South Korea using machine learning methods. Forest fires triggered by training cause considerable damage, but the number of occurrences is relatively very small, making it difficult to construct a general prediction model. We consider the forest fire prediction problem, which often occurs during the artillery fire training, as a classification of imbalanced data and propose the following two-step method. First, the case of forest fires is set as a minority class, and the majority class and the minority class are artificially evenly matched by employing an over-sampling technique based on a probability distribution of the minority data. In the second step, we use a weighted SVM that enforces a misclassification cost of minority class. The contribution of our research is as follows. This study contributes to significantly reducing the probability of forest fires during the artillery military training of the ROKA. It can also lessen forest fire catastrophes, while simultaneously considering reducing the numerous adverse effects that the Korean Army should have endured due to forest fires caused by artillery military training. Second, it contributes to encouraging the development of practical approaches for forest fire prediction in regions with a climate similar

Table 1. Performance comparison for linear SVM, Kernel SVM and GC-WSVM.

metric	GC-WSVM	Linear SVM	Kernel SVM
g-mean	0.891 (0.017)	0 (NaN)	0.148 (0.097)
sensitivity	0.886 (0.033)	0 (NaN)	0.079 (0.063)
specificity	0.901 (0.014)	1 (NaN)	0.997 (0.002)

The numbers in parentheses are standard errors.

to that of the Korean Peninsula. The Korean Peninsula has a moderate temperate climate zone that is suitable for research in various regions. Finally, our proposed method can facilitate the study of classification models for various imbalanced data.

RESULT

This section presents and interprets the result of applying the proposed method to the artillery firing training dataset which is described in the method section. We compared linear support vector machine (linear SVM), kernel support vector machine (kernel SVM) and Gaussian mixture Clustering Weighted SVM (GC-WSVM), which will be described in detail in the next section. We have used R packages for experiments; *mclust* for clustering using Gaussian mixture model (GMC), *mvtnorm* for multivariate normal random sample generation, *e1071* for linear and kernel SVM, *WeightSVM* for weighted SVM, and *caret* for 5-fold cross validation. All experiments were conducted with 2.8 GHz Intel Core i7 quad-core processor (4558U) and 16GB RAM.

G-mean, which is the geometric mean of specificity and sensitivity, was used as the metric for the performance of the classifying algorithm. Since the specificity represents the ability to capture the majority class and the sensitivity represents the ability to identify the minority class, the g-mean successfully evaluates whether the algorithm has balanced classification performance for both classes. Instead of separating the original dataset into two parts, namely the training dataset and test dataset, we adopted 5-fold cross-validation for evaluating the g-mean. This was because the number of samples in the minority class was too small, so the complete exclusion of the whole test dataset from the training process would have resulted in poor training of the algorithm. In each fold, 80% of the dataset was used for training and 20% was used for g-mean evaluation. Within that 80% of the training dataset, we again conducted 5-fold cross validation for hyperparameter tuning. We repeated one hundred times of the procedure explained above, which resulted in one hundred different splittings. This Monte Carlo simulation was conducted in purpose of reducing the volatility produced by random splitting of 5-fold cross validation when calculating the performance metric, thus helping us to check the performance more clearly. This process also allows us to see the standard error of the algorithms, which represents the generalization ability of the algorithms to the new dataset. We present the mean and standard deviation of the resulting g-means, specificities and sensitivities.

Table 1 shows the performance comparison results for all methodologies. In the results, we can see that the proposed method outperforms competitors. Specifically, the proposed method obtains a high g-mean value with uniformly high sensitivity and specificity. Linear SVM completely fails to predict the minority class, and results in 1 specificity and 0 sensitivity, which leads to 0 g-mean. Kernel SVM produced slightly better sensitivity and g-mean than the linear SVM, but its overall performance falls short of GC-WSVM.

We first tried the conventional linear SVM method. Despite its overly simple assumptions of linear separability, linear SVM is widely used for the quick prediction of a massive dataset. Since in highly imbalanced datasets there are dozens or hundreds of times more negative instances than positive instances, the size of the dataset is usually large. SVMs need to be fitted hundreds of times because of hyperparameter tuning. Thus linear SVM, which has $O(n^2)$ time complexity in its modern implementation, has an advantage in this sense, compared to kernel SVMs whose time complexity is $O(n^3)$. Linear SVM has one hyperparameter, C, which represents the overall penalty for misclassification(former definition and formulation of SVM is explained in detail in the Method seciton). This hyperparameter was chosen from the range of $2^{-10} - 2^{10}$ by 5-fold cross validation, using g-mean as the performance metric. However, the performance of the linear SVM classifier is not so poor that it cannot be used in practice. We observed a sensitivity of 0 and a specificity of 1, meaning that the linear SVM outputted non-wildfire for every instance. This behavior results in perfectly predicting all non-wildfire while non of the wildfire cases are identified. In other words, the linear SVM produces meaningless results in imbalanced classification. As stated before, each class in a classification problem usually consists of subgroups, and in many cases, these subgroups are not clustered among themselves, but rather spread out in feature space. This leads to a nonlinear distribution of each class, and with the additional class imbalance issue, the linear SVM completely fails to find the right decision boundary. We next evaluated kernel SVM with Gaussian kernel. The method gained its popularity due to its nice performance in nonlinear classification in various types of datasets. Gaussian kernel is considered almost de facto when using the kernel SVM, in both research and practical data analysis.

Table 2. Meteorological variables provided by South Korea Meteorological Administration

Meteorological variable	Detailed Variable Description	Unit
Temperature	Temperature at 2m	Degree Celsius
Precipitation	Accumulative precipitation over last 60 minutes	mm
Wind speed	Average of 10 minute at 10m altitude	m/s
Wind direction	Average of 10 minute at 10m altitude	Degree
Relative humidity	-	%

Gaussian kernel requires additional hyperparmeter γ, which is the bandwidth of the kernel that represents the extent to which the influence of each sample can be reached. As with linear SVM, the hyperparemters were tuned by 5-fold cross-validation, with the range of γ being $2^{-10} - 2^{10}$, using g-mean as the performance metric. The results were disappointing with g-mean of 0.148. Although the specificity was very high (0.997), the sensitivity was very low (0.079). So the nonlinear classifying ability granted by Gaussian kernel has only slightly increased the chance of capturing wildfire cases. This implies that the kernel SVM is not enough for highly imbalanced classification, and some other tricks need to be applied. Finally, we applied our GC-WSVM method. This method first creates synthetic minority instances considering the estimated distribution of the minority class by the GMC method. Then it specifies different costs of misclassifying for each class when applying the kernel SVM. This resulted in the g-mean of 0.891, with the sensitivity of 0.886 and the specificity of 0.901. This result is meaningful in that sensitivity and specificity are balanced, since this implies that our proposed algorithm successfully predicts both of wildfire and non-wildfire cases. Weights were specified using the method explained in Method section. C and γ values were tuned in the same way as the kernel SVM. The original dataset has the imbalance ratio of 24.6. In our method we treat the oversampling ratio as a hyperparameter; we tried six oversampling ratio, each of which resulting in the imbalance ratio of $20, 15, 10, 5, 2$ and 1. For each imbalance ratio, we separately conducted the procedure described above. With the hyperparameters properly chosen, using 5-fold cross-validation on the imbalance ratio with g-mean metric, we selected the optimal imbalance ratio. The oversampling ratio resulting in the imbalance ratio of 15 was chosen. It should be noted that despite the random nature of oversampling, the GC-WSVM has a lower standard error of sensitivity and g-mean than kernel SVM. Since there are only a few minority training samples in the highly imbalanced dataset, random splitting by 5-fold cross-validation greatly changes the distribution of the training dataset, resulting in high standard error. GMC SMOTE recovers the distribution of the minority class, filling in the blank. Consequently, it results in a more stable performance when predicting the minority class.

CONCLUSION

Accurate forecasts of wildfires prevent devastating effects on the economy, save lives and prevent ecological damage. In this work, we proposed machine learning approach to predict forest fire risk for Artillery training in South Korea. We viewed the task as a classification of imbalanced data and developed a two-phased method. First, we alleviate the class imbalance by oversampling the minority class based on a probability distribution. Second, we apply weighted SVM to the balanced dataset. The proposed method outperforms classical classification methods such as linear SVM and kernel SVM by large margins.

Dozens of divisions and artillery brigades of the ROKA are very qualitative and subjective in making decisions concerning artillery military training, which can occasionally result in a relatively high risk of forest fires or military training that is below the necessary level. This study can contribute to reasonably and objectively providing guidelines on whether artillery military training is conducted. In conclusion, this not only significantly reduces the probability of forest fires during conventional artillery fire training in the ROKA but also encourages the development of practical approaches for wildfires prediction in countries with climates similar to the Korean Peninsula. Furthermore, our open-sourced dataset and the proposed method can contribute to the study of the imbalanced data and design of various classification models.

METHOD

Database

Based on weather observation data from the South Korea Meteorological Administration (SKMA), source data were collected every hour from 612 observation stations nationwide starting from January 1 2011 to Jan 31 2021. There are two types of stations, namely Automated Synoptic Observing System (ASOS) and Automatic Weather System (AWS). Observation stations provide meteorological variables such as temperature, precipitation, wind, and humidity. Geographical location of the stations are given by elevation, latitude, and longitude, and they are used later to normalize the meteorological variables. The temperature

Table 3. Sample forest fires reported during the artillery firing training

No	Date	Location	Shooting type	Damaged Area
1	31.Oct.2019, 20:44	Goseong-gun, Gangwon-do	Red parachute flare	2,500
2	8. Nov.2019, 12:15	Goseong-gun, Gangwon-do	Star shell	3,000
3	13.Apr.2017, 20:10	Paju-si, Gyeonggi-do	60mm trench mortar	150,000
4	22.Mar.2018, 13:51	Paju-si, Gyeonggi-do	Panzerfaust3	3,300

and humidity data are available at various time intervals: 1 minute, 1 hour, and 1 day. The temperature was expressed as 2m high instantaneous temperature in degrees Celsius. The wind uses a 10-minute average value, and the wind direction starts at 0 degrees in the north and is expressed at 360 degrees. For precipitation, cumulative precipitation for 1 hour was used, and for humidity, relative humidity at that time was used. Table 2 describes meteorological variables provided by SKMA.

Observation stations are not uniformly distributed across nationwide and, most of the time, there are few to none stations nearby the artillery training sites. Hence, measurements provided by the stations are not accurate representation of the weather conditions of the artillery training sites. To get more precise characteristics of the site, first we apply Barnes interpolation [46] to get uniformly gridified map of weather measurements at 1km resolution. Next, we use high resolution (30m) Digital Elevation Model (DEM) of the training sites provided by Shuttle Radar Topography Mission (SRTM) to normalize the meteorological variables based on elevations. The temperature was increased from the observation station altitude to 0m according to the lapse rate (0.65/100m), and then gridified to decrease again to the altitude of the 30m high-resolution terrain elevation map.

Dataset

Over the past five years, about 996 artillery training sessions have been reported, of which only 40 have resulted in forest fires. Of the 40 reported wildfires, 39 occurred in the February-April and October-November periods, with only one in June. Table 3 shows 4 samples that reported forest fire cases. Note that, the remaining 36 cases were not included in the Table 3 due to the space limitation, but it can be provided upon request. Each training event reports date, time, location, the shooting type and total damaged area. We build a artillery training forest fire dataset by merging time and location information of the training sessions with meteorological variable database provided by the SKMA. Each training session has four feature variables: precipitation, temperature, wind speed, and humidity. We recall that our dataset contains 1036 samples (996 no wildfire and 40 wildfire occurrence), and the data imbalance ratio is around 1:25.

As can be seen in figure 1, meteorological variables have different distribution ranges. Temperature varies from 20 to 40, precipitation is 0 to 80, wind speed is 0 to 20, and humidity is 0 to 100. Uniform weights cannot be used during training because the distribution of each value is different. Therefore, variables were normalized to have a zero mean and a standard deviation of one. Some temperature values were negative and thus have been converted to absolute temperature. The normalization parameters are later used again during the validation and testing. From Figure 1, it can be intuitively seen that each of the four variables has a significant difference in the presence or absence of forest fires. Next, in figure 2 each axis of the the scatter plot represents humidity, temperature, and wind speed, and the size of the point denotes precipitation. Also, the size of point denotes precipitation. It is worth noting that a small number of forest fire cases form several clusters, suggesting that a probability distribution can be derived.

In figure 3 the four-dimensional features of the our data is reduced to two-dimensions using a classical multidimensional scaling (MDS) technique [47], which is also known as Principal Coordinates Analysis (PCoA), Torgerson Scaling or Torgerson–Gower scaling. Although it is not rigorous because the dimension has been reduced, from the location of the red dots representing 40 forest fire events, it would be reasonable to assume that minority class are generated from multiple clusters rather than from one probability distribution.

The collected dataset is rare and valuable, and we believe it can be used for a variety of other future research and applications, such as developing practical approaches for forest fire prediction in countries with climate similar to the Korean Peninsula, and developing classification algorithms for highly imbalanced data. The dataset and the source code have been approved by the Republic of Korea Military Security and thus will be publicly available.

Machine Learning Approach

We take a two-step approach as the modeling strategy. Specifically, we first alleviate the class imbalance by oversampling the minority class, and then we apply *WeightSVM* to the balanced dataset. We assume that the minority class follows a Gaussian mixture distribution for the oversampling. This assumption means that data points are divided into several small groups, and

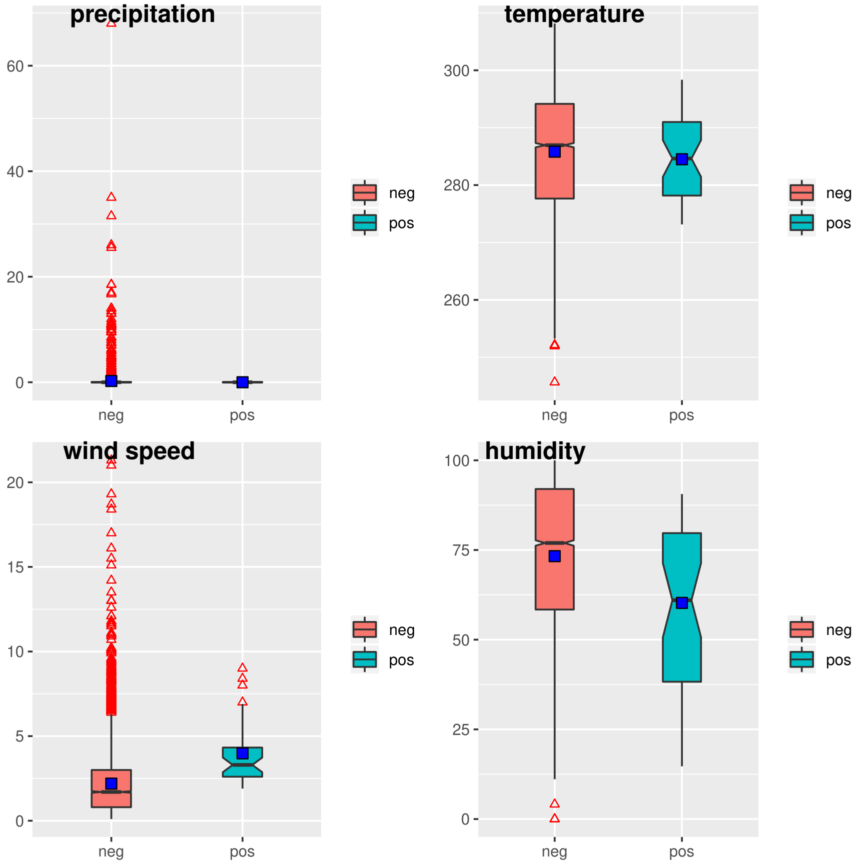

Figure 1. Meteorological feature distributions for positive and negative cases. Note that negative cases represent wildfire outbreaks. Positive cases denote absence of wildfire outbreaks.

each small group follows a normal distribution. We estimate the parameters of the Gaussian mixture distribution and generate new instances from that distribution. In many cases, the assumption is consistent with reality. Rare events, despite the small number of their occurrences, are often divided into several small subgroups. This is because in some cases, especially when there are not many predictors, different values of predictors may produce the same result. In many cases of data analysis, we do not have enough predictors to fully explain the data-generating process. In the cases, even if the underlying data-generating process does not have clusters, the lack of predictors in data analysis leads to the clustering of rare events. Even if the minority class does not consist of subgroups, we can approximate most continuous distributions with Gaussian mixture distributions, in the sense that any smooth density can be approximated with any specific nonzero amount of error by the Gaussian mixture distribution with enough components [48]. SMOTE algorithm [25], which is widely adopted due to its simplicity, only creates new instances from a straight line between existing instances. Thus it does not effectively broaden the area of the minority class. Also, if the minority class consists of subgroups, SMOTE could incorrectly generate new instances in the area between two subgroups.

Now we briefly explain the Gaussian mixture model (GMM). Let $f(x)$ be the probability density function of the minority class. The GMM assumes that $f(x)$ is represented as a linear combination of K Gaussian density functions as follows:

$$f(x) = \sum_{k=1}^{K} \pi_k \mathscr{N}(x|\mu_k, \Sigma_k),$$

where $\mathscr{N}(\cdot \mid \mu_k, \Sigma_k)$ denotes a Gaussian distribution with mean vector μ_k and covariance matrix Σ_k, and π_k are mixing coefficients satisfying the constraints that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Note that the mixing coefficients indicate how much of each Gaussian density contributes to the total distribution. The parameters $\{\mu_k, \Sigma_k, \pi_k\}$ of the GMM are estimated by maximizing the log-likelihood $\mathscr{L}(\pi_k, \mu_k, \Sigma_k | x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi_k \mathscr{N}(x_i \mid \mu_k, \Sigma_k))$ using the expectation-maximization (EM) algorithm [49], which is an algorithm to find the maximum likelihood estimates iteratively. Here n denotes the number of training observations.

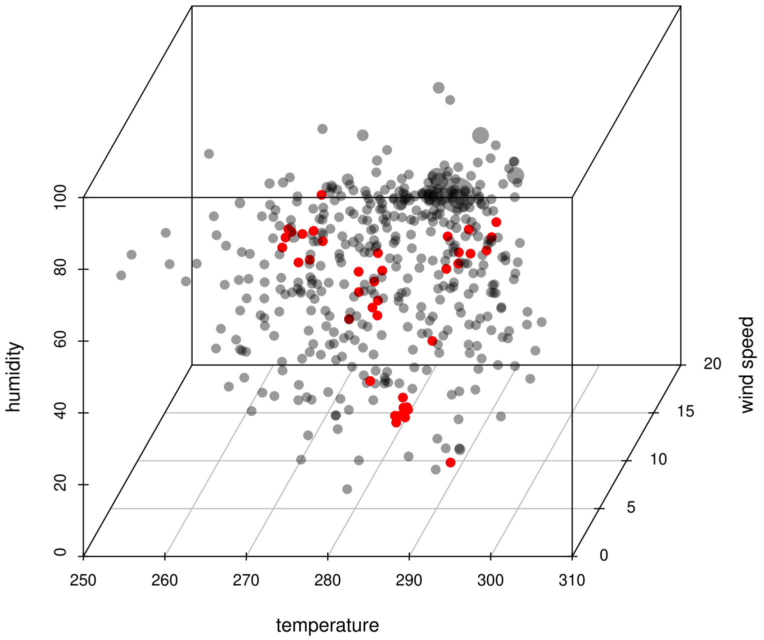

Figure 2. Each axis of the three-dimensional graph represents humidity, temperature, and wind speed, and the size of the point denotes precipitation. Red dots are 40 events in which forest fires have occurred, and gray dots are marked by selecting 996 random events.

In the second step, we apply *WeightSVM*, which means we put more weight on the cost of misclassifying the minority class (wildfire) compared to the misclassification cost of majority class (absence of wildfire). This is in line with the reality, since a single wildfire in an unprepared situation costs much more than several preparations for wildfires that did not eventually occur. The misclassification costs are set proportional to the inverse of the number of instances. We briefly provide the formal definition of *WeightSVM*. It is a quadratic optimization problem of the form:

$$\min_{w,b,\xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} c_i\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i(w^T\phi(x_i)+b) \geq 1-\xi_i, \tag{2}$$

$$\xi_i \geq 0 \quad \text{for } i = 1,\ldots,n$$

Here x_i's and y_i's indicate predictors and label of each training observation, respectively, $\phi(\cdot)$ is a nonlinear feature mapping that gives nonlinearity to the decision boundary of the SVM, and the squared euclidean norm $||w||^2$ indicates the inverse of "margin", which is the distance between the decision boundary and each class. The SVM seeks to maximize the margin, to achieve small generalization error for future data. c_i's determine misclassification cost. We set $c_i = 1/n_{maj}$ for wildfire instances and $c_i = 1/n_{min}$ for non-wildfire instances, where n_{maj} and n_{min} indicate the number of instances in the majority class and the minority class, respectively. C controls the total penalty level to misclassification. The nonlinear mapping $\phi(\cdot)$ may be infinite-dimensional, which gives great classification ability to the SVM. Infinite-dimensional feature mapping is only feasible to calculate when the objective function above is solved by transforming it into its Lagrangian dual form given by

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{3}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C. \tag{4}$$

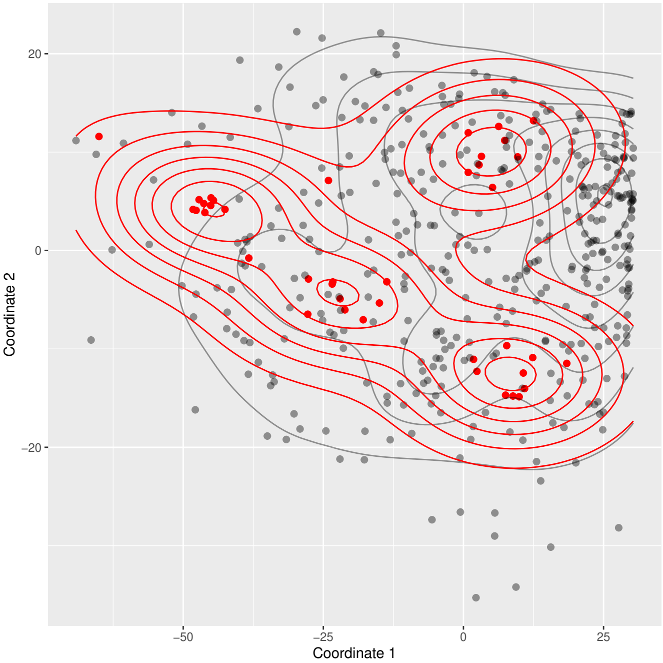

Figure 3. The dimension of data is reduced to two dimensions using a multidimensional scaling technique. Red dots are 40 events in which forest fires have occurred, and gray dots represent the absence of wildfire outbreaks

Table 4. A confusion matrix for imbalanced problem

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

where α_i's are the Lagrangian multipliers and $K(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function that gives the inner product of the feature mapping of x_i and x_j. The solution α_i^* of the above problem gives the separating hyperplane $g(x) = \sum_{i=1}^{N} \alpha_i^* y_i K(x_i,x_j) + b$. The SVM classifies classifies the input vector x as positive instance (wildfire) if $g(x) > 0$.

Performance metric for imbalanced data

When it comes to imbalanced data classification performance metric, simple error rate is not appropriate, because we can obtain small error rate just by classifying all instances as the majority class. Therefore typically a confusion matrix (Table 4) is used to measure the performance of the classifier. In the confusion matrix, minority class has positive label and the majority class has negative label. Using confusion matrix (Table 4) we define *Accuracy*, *Sensitivity*, *Specificity*, and *G-mean* as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP+TN} \tag{6}$$

$$Specificity = \frac{TN}{FN+FP} \tag{7}$$

$$G-mean = \sqrt{Sensitivity \times Specificity} \tag{8}$$

The specificity measures the ability to capture the majority class and the sensitivity represents the ability to identify the minority class. G-mean is the geometric mean of specificity and sensitivity, which evaluates whether the algorithm has balanced classification performance for both classes.

Acknowledgements

Jaeoh Kim was supported by INHA University Research grant. Seongil Jo was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1A01013338).

Author contributions statement

N.J., S.J., and J.K. conceived and designed the study. J.K. performed the data management and collection. All analyses were performed by N.J., S.J., and J.K. This article was drafted by N.J., and J.M. All authors reviewed the manuscript before submission.

Data and code availability

Data and source code underlying the results presented in this paper can be obtained from the authors upon reasonable request.

Disclosures

The authors declare no competing interests.

References

1. FOGARITI, L. & CATCHPOLE, P. I. W. Adoption vs. adaptation: Lessons from applying the canadian forest fire danger rating system in new ze. .

2. López, A. S., San-Miguel-Ayanz, J. & Burgan, R. E. Integration of satellite sensor data, fuel type maps and meteorological observations for evaluation of forest fire risk at the pan-european scale. *Int. J. Remote. Sens.* **23**, 2713–2719 (2002).

3. Van Wagner, C., Forest, P. *et al.* Development and structure of the canadian forest fireweather index system. In *Can. For. Serv., Forestry Tech. Rep* (Citeseer, 1987).

4. Stocks, B. J. *et al.* The canadian forest fire danger rating system: an overview. *The For. Chron.* **65**, 450–457 (1989).

5. FARSITE, F. M. Fire area simulator-model development and evaluation. *Res. Pap. RMRS-RP-4 Revised. Ogden, UT: USDA For. Serv. Rocky Mountain Res. Stn.* (2004).

6. Kloprogge, P., Van der Sluijs, J. P. & Petersen, A. C. A method for the analysis of assumptions in model-based environmental assessments. *Environ. Model. & Softw.* **26**, 289–301 (2011).

7. Yu, Y. *et al.* Machine learning–based observation-constrained projections reveal elevated global socioeconomic risks from wildfire. *Nat. communications* **13**, 1–11 (2022).

8. Crowley, G. *et al.* Assessing the protective metabolome using machine learning in world trade center particulate exposed firefighters at risk for lung injury. *Sci. reports* **9**, 1–10 (2019).

9. Thach, N. N. *et al.* Spatial pattern assessment of tropical forest fire danger at thuan chau area (vietnam) using gis-based advanced machine learning algorithms: A comparative study. *Ecol. informatics* **46**, 74–85 (2018).

10. Jafari Goldarag, Y., Mohammadzadeh, A. & Ardakani, A. Fire risk assessment using neural network and logistic regression. *J. Indian Soc. Remote. Sens.* **44**, 885–894 (2016).

11. Rodrigues, M. & De la Riva, J. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ. Model. & Softw.* **57**, 192–201 (2014).

12. Al-Fugara, A. *et al.* Wildland fire susceptibility mapping using support vector regression and adaptive neuro-fuzzy inference system-based whale optimization algorithm and simulated annealing. *ISPRS Int. J. Geo-Information* **10**, 382 (2021).

13. Kim, S., Lee, W., Park, Y.-s., Lee, H.-W. & Lee, Y.-T. Forest fire monitoring system based on aerial image. In *2016 3rd international conference on information and communication technologies for disaster management (ICT-DM)*, 1–6 (IEEE, 2016).

14. Jiao, Z. *et al.* A deep learning based forest fire detection approach using uav and yolov3. In *2019 1st International conference on industrial artificial intelligence (IAI)*, 1–5 (IEEE, 2019).

15. Xu, R., Lin, H., Lu, K., Cao, L. & Liu, Y. A forest fire detection system based on ensemble learning. *Forests* **12**, 217 (2021).

16. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge data engineering* **21**, 1263–1284 (2009).

17. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intell. data analysis* **6**, 429–449 (2002).

18. Debnath, T. & Nakamoto, T. Predicting individual perceptual scent impression from imbalanced dataset using mass spectrum of odorant molecules. *Sci. reports* **12**, 1–9 (2022).

19. Sun, Y., Wong, A. K. & Kamel, M. S. Classification of imbalanced data: A review. *Int. journal pattern recognition artificial intelligence* **23**, 687–719 (2009).

20. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232 (2016).

21. Haixiang, G. *et al.* Learning from class-imbalanced data: Review of methods and applications. *Expert. systems with applications* **73**, 220–239 (2017).

22. Tomek, I. Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.* **6**, 769–772 (1976).

23. Zhao, X.-M., Li, X., Chen, L. & Aihara, K. Protein classification with imbalanced data. *Proteins: Struct. function, bioinformatics* **70**, 1125–1132 (2008).

24. Khalilia, M., Chakraborty, S. & Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics decision making* **11**, 1–13 (2011).

25. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. artificial intelligence research* **16**, 321–357 (2002).

26. Sun, Y., Kamel, M. S., Wong, A. K. & Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition* **40**, 3358–3378 (2007).

27. Bunkhumpornpat, C., Sinapiromsaran, K. & Lursinsap, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, 475–482 (Springer, 2009).

28. Bekkar, M., Djemaa, H. K. & Alitouche, T. A. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl* **3** (2013).

29. Kubat, M., Matwin, S. *et al.* Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, vol. 97, 179 (Citeseer, 1997).

30. Nguyen, H. M., Cooper, E. W. & Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigms* **3**, 4–21 (2011).

31. Ramentol, E. *et al.* Smote-frst: a new resampling method using fuzzy rough set theory. In *Uncertainty modeling in knowledge engineering and decision making*, 800–805 (World Scientific, 2012).

32. López, V., Fernández, A., García, S., Palade, V. & Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. sciences* **250**, 113–141 (2013).

33. Anand, R., Mehrotra, K. G., Mohan, C. K. & Ranka, S. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks* **4**, 962–969 (1993).

34. Chawla, N. V., Japkowicz, N. & Kotcz, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter* **6**, 1–6 (2004).

35. Han, H., Wang, W.-Y. & Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887 (Springer, 2005).

36. Mani, I. & Zhang, I. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, 1–7 (ICML, 2003).

37. Beckmann, M., Ebecken, N. F., de Lima, B. S. P. *et al.* A knn undersampling approach for data balancing. *J. Intell. Learn. Syst. Appl.* **7**, 104 (2015).

38. Barandela, R., Valdovinos, R. M., Sánchez, J. S. & Ferri, F. J. The imbalanced training sample problem: Under or over sampling? In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, 806–814 (Springer, 2004).

39. Turney, P. D. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *J. artificial intelligence research* **2**, 369–409 (1994).

40. Ling, C. X., Yang, Q., Wang, J. & Zhang, S. Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, 69 (2004).

41. Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Syst. Man, Cybern. Part B (Cybernetics)* **39**, 539–550 (2008).

42. Gao, M., Hong, X., Chen, S. & Harris, C. J. A combined smote and pso based rbf classifier for two-class imbalanced problems. *Neurocomputing* **74**, 3456–3466 (2011).

43. Fernández, A., del Jesus, M. J. & Herrera, F. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *Int. J. Approx. Reason.* **50**, 561–577 (2009).

44. Chawla, N. V., Lazarevic, A., Hall, L. O. & Bowyer, K. W. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, 107–119 (Springer, 2003).

45. Mease, D., Wyner, A. J. & Buja, A. Boosted classification trees and class probability/quantile estimation. *J. Mach. Learn. Res.* **8** (2007).

46. Barnes, S. L. A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteorol. Climatol.* **3**, 396–409 (1964).

47. Cox, M. A. & Cox, T. F. Multidimensional scaling. In *Handbook of data visualization*, 315–347 (Springer, 2008).

48. Goodfellow, I. J., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).

49. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc. Ser. B (Methodological)* **39**, 1–22 (1977).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PredictionofForestFireRiskforArtilleryTrainingusingWeightedSupportVectorMachineforimbalanceddata.zip