applied sciences

MDPI

Article

Exploiting Domain Knowledge to Address Class Imbalance in Meteorological Data Mining

Evangelos Tsagalidis [1] and Georgios Evangelidis [2,*]

[1] Hellenic Agricultural Insurance Organization, Meteorological Applications Centre, International Airport Makedonia, 551 03 Thessaloniki, Greece
[2] Department of Applied Informatics, School of Information Sciences, University of Macedonia, 546 36 Thessaloniki, Greece
* Correspondence: gevan@uom.edu.gr

Abstract: We deal with the problem of class imbalance in data mining and machine learning classification algorithms. This is the case where some of the class labels are represented by a small number of examples in the training dataset compared to the rest of the class labels. Usually, those minority class labels are the most important ones, implying that classifiers should primarily perform well on predicting those labels. This is a well-studied problem and various strategies that use sampling methods are used to balance the representation of the labels in the training dataset and improve classifier performance. We explore whether expert knowledge in the field of Meteorology can enhance the quality of the training dataset when treated by pre-processing sampling strategies. We propose four new sampling strategies based on our expertise on the data domain and we compare their effectiveness against the established sampling strategies used in the literature. It turns out that our sampling strategies, which take advantage of expert knowledge from the data domain, achieve class balancing that improves the performance of most classifiers.

Keywords: meteorological data mining and machine learning; class imbalance; classification; randomized undersampling; SMOTE oversampling; undersampling using temporal distances

check for updates

Citation: Tsagalidis, E.; Evangelidis, G. Exploiting Domain Knowledge to Address Class Imbalance in Meteorological Data Mining. *Appl. Sci.* **2022**, *12*, 12402. https://doi.org/10.3390/app122312402

Academic Editor: Yosoon Choi

Received: 27 October 2022
Accepted: 2 December 2022
Published: 4 December 2022

1. Introduction

Imbalanced or skewed training datasets make predictive modeling challenging since most of the classifiers are designed assuming a uniform distribution of class labels among the examples. There are classification problems that must deal with various degrees of imbalance. The goal is to improve the quality of the training dataset, i.e., make it more balanced, in order for the classifiers to achieve better predictive performance, specifically for the minority class. Usually, the minority class is more important and, hence, the classifier should be more sensitive to classification errors for the minority class than the majority class [1]. A typical approach in the literature is the application of techniques for transforming the training dataset to balance the class distribution including data oversampling for the minority examples, data undersampling for the majority examples and combinations of these techniques [1,2].

We attempt to enhance existing pre-processing sampling strategies by exploiting expert knowledge from the domain of Meteorology. We use the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis 40-years dataset (See https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-40-years, accessed on 29 November 2022, for details) (also known as ERA-40) and a dataset with the historical observations of the meteorological station of Micra, Thessaloniki, Greece and we attempt to predict the occurrence of precipitation on the ground at the meteorological station. We use various data pre-processing strategies (based on oversampling and undersampling) for the selection of the appropriate training dataset, and, we test their effectiveness on various classifiers.

The input dataset consists of imbalanced data regarding the precipitation class variable, where the minority class is only the 16.1% of the cases. It is known that such situations degrade the performance of data mining or machine learning classifiers. In [3], we determined the minimum training dataset size that can ensure effective application of data mining techniques specifically on meteorological data. The performance of various classifiers did not increase significantly for training dataset sizes of more than 9 years. Also, the results were not affected by the way we chose the training dataset examples, i.e., randomly isolated examples totalling nine years versus nine entire yearly sets of examples randomly selected. In this paper, we take advantage of the above finding by choosing appropriately large training datasets for the tested classifiers.

The contribution of this study is the proposal of effective sampling strategies on meteorological training datasets that are based on our expertise on the data domain. In our experimental study, we compare common sampling strategies from the literature and the proposed new strategies and show that the newly proposed strategies improve the performance of most classifiers.

The remainder of the paper is organized as follows. Section 2 discusses the problem of class imbalance, reviews recent works that address it using domain knowledge, and, describes the sampling strategies used in the literature as well as the novel sampling strategies we propose. Section 3 describes the datasets we used for applying the sampling strategies on the training dataset and the classifiers that we compared. Section 4 discusses the methodology used in the experiments. In Section 5, we present the analysis and the results, and, finally, we conclude in Section 6.

2. The Problem of Class Imbalance

A very good introduction to the problem of class imbalance and the related research efforts is given in [4,5]. Ref. [4] provides a comprehensive review of the subject and discusses the initial solutions that were proposed to deal with the problem of class imbalance. Ref. [5] discusses the role that rare classes and rare cases play in data mining, the problems that they can cause and the methods that have been proposed to address these problems.

Over the years the problem of class imbalanced has been studied extensively. There exist numerous papers that use standard data agnostic oversampling and undersampling techniques to create balanced training datasets. Regarding meteorological data, ref. [6] first applies oversampling to increase thunderstorm examples in the training dataset and then uses deep neural networks to predict thunderstorms. Similarly, ref. [7] applies standard oversampling techniques on radar image data to improve rainfall prediction, while [8] presents a framework for predicting floods, in which it embeds re-sampling to address class imbalance. Finally, ref. [9] does not apply any sampling strategies but experiments with various classifiers and concludes that Self-Growing Neural Networks perform better when predicting fog events using data with class imbalance.

Various research works attempt to exploit domain knowledge to address the class imbalance problem, but not in the meteorological domain. Ref. [10] addresses the problem of noisy and borderline examples when using oversampling methods, while [11] deals simultaneously with the problems of class imbalance and class overlap. Ref. [12] uses domain specific knowledge to address the problem of class imbalance in text sentiment classification. Finally, ref. [13] exploits domain knowledge to address multi-class imbalance in classification tasks for manufacturing data.

In our study, we use the most common sampling strategies found in the literature to address the class imbalance problem, namely, the randomized undersampling and the SMOTE oversampling methods and their combination. SMOTE stands for Synthetic Minority Oversampling Technique [14]. Besides the natural distribution, we employ the commonly used 30% and 50% (or balanced) distributions regarding the minority class [1]. We also examine the within-class distribution in addition to the between-class distribution [15], using a combination of the randomized undersampling and the SMOTE oversampling methods in both minority and majority examples.

In an effort to take into account the peculiarities of the data domain when sampling the training datasets and to examine how these could affect the performance of the classifiers, we applied two novel strategies when constructing balanced datasets, i.e., datasets where the number of majority and minority examples is equal. In the first strategy, we applied the k-Means clustering algorithm using "classes to clusters" evaluation to select only the most homogeneous majority examples. In the second strategy, we rejected the majority examples that were closer to the minority examples with respect to their temporal distance in days using three different values for the distance. Then, we further reduced the number of majority examples to achieve a balanced distribution using the randomized undersampling method. We are not aware of any other attempt that uses large meteorological databases and at the same time domain specific sampling techniques to address the class imbalance problem.

We used five different classifiers to build models for predicting our class variable. The training/test set method was used to evaluate the models and to reveal the best sampling strategy for meteorological data. As an evaluation metric, we used the Area Under the ROC (Receiver Operating Characteristics) Curve (AUC) [5,16].

3. Datasets

3.1. ERA-40 Dataset

The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis 40-years dataset (ERA-40) is a global atmospheric analysis of many conventional observations and satellite data streams for the period of September 1957 to August 2002. Reanalysis products are used increasingly in many fields that require an observational record of the state of either the atmosphere or its underlying land and ocean surfaces. There are numerous data products that are separated into dataset series based on resolution, vertical coordinate reference, and likely research applications. In this study, we used the ERA-40 2.5 degree latitude-longitude gridded upper air analysis on pressure surfaces. This dataset contains 11 variables on 23 pressure surfaces on an equally spaced global 2.5 degree latitude-longitude grid. All variables are reported four times a day at 00, 06, 12 and 18UTC for the entire period [17].

We created our initial dataset choosing the values of 10 variables on 7 pressure surfaces on one node. We used only the data from the node with geographical coordinates 40° N latitude and 22.5° E longitude, which is the closest node to the Meteorological Station of Micra, Thessaloniki, Greece located at 40.52° N, 22.97° E and altitude of 4m. We omitted the 11th variable of the Ozone mass mixing ratio. The 1000 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa, 300 hPa and 200 hPa are the 7 pressure surfaces we chose, because these are the ones that are mainly used by the meteorology forecasters operationally. In addition, the values of the barometric pressure on mean sea level in Pa supplement the initial dataset that consists of 71 variables.

Furthermore, the initial values of most of the variables for each pressure surface and the pressure on mean sea level were transformed to make them easier to understand or to express them in the same metric units as used operationally by the meteorologists. More specifically, specific humidity initially expressed in $kg \cdot kg^{-1}$ was converted to $g \cdot kg^{-1}$ and vertical velocity in $Pa \cdot s^{-1}$ to $hPa \cdot h^{-1}$. The relatively small values of both vorticity (relative) in s^{-1} and divergence also in s^{-1} were multiplied by 10^6, and the value of potential vorticity in $K \cdot m^2 \cdot kg^{-1} \cdot s^{-1}$ by 10^8. Regarding the wind, wind direction in azimuth degrees and wind speed in knots were calculated using the U and V velocities in m^{-1}. Also, the azimuth degrees for the wind direction were assigned into the eight discrete values of north (N), northeast (NE), etc., used in meteorology. The geopotential in $m^2 \cdot s^{-2}$ was divided by the World Meteorological Organization (WMO) defined gravity constant of 9.80665 $m \cdot s^{-2}$, thus, it was transformed to geopotential height in gpm. Finally, the values of barometric pressure on mean sea level were expressed in hPa, and only the values of temperature in K and relative humidity as percentage (%) on pressure surfaces remained unchanged.

3.2. Class Variable

The 6-hourly main synoptic surface observation data of the Meteorological Station of Micra, Thessaloniki, Greece completed our initial dataset. More specifically, we collected the recorded precipitation data of the period 1 January 1960 00UTC–31 December 2001 18UTC. We assigned the value 'yes' to the 6-hourly records of rain, drizzle, sleet, snow, shower at the station or the records of thunderstorm at the station or around it, and the value 'no' to the rest of the records, thus, creating the class variable of our study. Our purpose is to use the ERA-40 atmospheric analysis data at node 40° N, 22.5° E to predict the precipitation at the station. We mention that the determination of the recorded precipitation is taking into account both the present and past weather of the synoptic observation, and that snow or thunder have priority over rain. Tables 1 and 2 depict the distribution of the precipitation types that had been recorded in the Meteorological Station according to the defined sub-clusters.

Table 1. Natural distribution of values within the minority class variable (precipitation 'yes').

Rain/Drizzle	Snow/Sleet	Thunder	Total 'Yes'
7154	547	2181	9882
11.66%	0.89%	3.55%	16.1%

Table 2. Natural distribution of values within the majority class variable (precipitation 'no').

Fog	Fair/Cloudy	Total 'No'
1395	50,087	51,482
2.27%	81.62%	83.9%

3.3. Predictor Variables

In the pre-processing phase we applied data reduction using the Principal Component Analysis (PCA) extraction method to remove highly correlated variables from the ERA-40 dataset. We used the SPSS statistical software package to process the entire ERA-40 dataset and to produce a new one that consisted of a reduced number of uncorrelated variables.

After applying PCA and examining the component matrix of loadings and the variable communalities, we deleted a total of 36 variables from our initial dataset that consisted of 71 variables. The component model was re-specified six times with a final outcome of 35 variables and 9 components with eigenvalues greater than 1. This is exactly the same methodology we used in a previous work of ours [18]. The analysis revealed the findings of Table 3.

Table 4 displays the variance explained by the rotated components and additionally the corresponding nine most highly correlated variables. The Total column gives the eigenvalue, or amount of variance in the original variables accounted for by each component. The % of Variance column gives the ratio of the variance accounted for by each component to the total variance in all of the variables (expressed as a percentage). The % Cumulative column gives the percentage of variance accounted for by the first 9 components.

The first nine rotated components explain nearly 85.2% of the variability in the original variables and it is possible to considerably reduce the complexity of the data set by using these components, with a 14.8% loss of information. As a result, we can reduce the size of the ERA-40 dataset by selecting the 9 most highly correlated variables with the 9 principal components [18,19]. These meteorological parameters could express the state of the troposphere where precipitation is created and reaches the ground. The reduced ERA-40 dataset with the 9 chosen variables, as predictors, and the precipitation, as class variable, comprised our experimental dataset with 61,364 examples. The size of the dataset is explained by the fact that we have four daily examples (one every 6 h) for a period of 42 years (42 × 365 × 4 = 61,320 examples plus 11 × 4 = 44 examples for the 11 extra leap year days of that period).

Table 3. Most highly correlated variables to the rotated components.

Component	Most Highly Correlated Initial Variable	Other Highly Correlated Initial Variables
1st	geopotential height on 200 hPa	geopotential height in the upper levels, the temperature almost in all levels, and the specific humidity in low levels of the atmosphere
2nd	relative vorticity on 1000 hPa	relative vorticity in low levels, the geopotential height on 925 hPa, and the pressure at mean sea level
3rd	wind direction on 300 hPa	wind direction in middle and upper levels
4th	wind speed on 300 hPa	wind speed in upper levels
5th	wind speed on 925 hPa	wind speed in low levels
6th	divergence on 300 hPa	vertical velocity in the upper levels
7th	temperature on 200 hPa	relative vorticity on 200 hPa
8th	potential vorticity on 500 hPa	relative vorticity on 500 hPa
9th	wind direction on 925 hPa	wind direction in low levels

Table 4. Variance explained by rotated components and the representative variables.

Component	Variable	Total	% of Variance	% Cumulative
1st	geopotential height 200 hPa	9.8	28	28
2nd	relative vorticity 1000 hPa	4.2	11.9	39.9
3rd	wind direction 300 hPa	2.9	8.4	48.3
4th	wind speed 300 hPa	2.6	7.5	55.7
5th	wind speed 925 hPa	2.4	7	62.7
6th	divergence 300 hPa	2.3	6.7	69.4
7th	temperature 200 hPa	2.1	6	75.4
8th	potential vorticity 500 hPa	1.8	5	80.4
9th	wind direction 925 hPa	1.7	4.8	85.2

4. Methodology

Since the focus of our study was to address the class imbalance problem, we used a number of sampling strategies in order to balance the training datasets used in the classification task.

Besides the training dataset with the natural distribution of the precipitation values that are shown in Tables 1 and 2 (Strategy 1), we created nine more balanced training datasets following different strategies (Strategies 2–10) (Table 5). Two of them followed the 30% distribution regarding the minority class variable and the other seven the balanced distribution (50%). In the following we describe Strategies 2 through 10.

In the second and third strategies, we used the randomized undersampling method to remove examples producing two datasets with a 30% (U30) and a 50% (U50) distribution of the minority class, respectively [5].

Likewise, in the fourth and fifth strategies, we used a combination of the SMOTE oversampling method to create new examples of the minority class and the randomized undersampling method to remove examples from the majority class, achieving a 30% (SU30) and a 50% (SU50) distribution of the minority class, respectively. We ran the SMOTE oversampling method in the WEKA environment, using 3 nearest neighbors [14,20,21].

In the sixth strategy (BW), we formed balanced datasets not only between-classes but also within-classes [15]. Thus, we employed the randomized undersampling method to reduce the number of the examples for the large clusters of 'Rain/Drizzle' and 'Fair/Cloudy' and the SMOTE oversampling method to increase the number of the examples for the

small clusters of 'Snow/Sleet', 'Thunder' and 'Fog'. Thus, the sum of the 'Rain/Drizzle', 'Snow/Sleet' and 'Thunder' examples that belong to the minority class became equal to the sum of the 'Fair/Cloudy' and 'Fog' examples of the majority class achieving the between-class balance. Moreover, the number of the 'Rain/Drizzle', 'Snow/Sleet' and 'Thunder' examples became equal to each other, and, similarly, the number of 'Fair/Cloudy' and 'Fog' examples became equal to each other achieving the within-class balance.

Table 5. Description of used sampling strategies.

Strategy	Acronym	Description
1	UN	Initial unbalanced dataset
2	U30	Randomized Undersampling 30%
3	U50	Randomized Undersampling 50%
4	SU30	SMOTE Oversampling + Randomized Undersampling 30%
5	SU50	SMOTE Oversampling + Randomized Undersampling 50%
6	BW	Balanced between-classes and within subclasses
7	CU	Remove majority examples that cluster with minority ones + Randomized Undersampling
8	D1U	Select only majority examples > 1 day away from minority ones + Randomized Undersampling
9	D2U	Select only majority examples > 2 days away from minority ones + Randomized Undersampling
10	D4U	Select only majority examples > 4 days away from minority ones

Strategies 7 through 10 are newly proposed sampling strategies that take into consideration the nature of the data at hand. More specifically, in the seventh strategy (CU), we applied the k-means clustering algorithm to the entire dataset using WEKA. We set the number of clusters equal to five and chose the "classes to clusters" evaluation in WEKA to evaluate each cluster according to the five classes of precipitation (Tables 1 and 2). In the first step, we selected only the majority examples of the 'Fair/Cloudy' and 'Fog' labeled clusters. In this manner, we rejected all the majority examples that clustered in the three clusters that corresponded to the three minority classes. The idea is that these examples are not good majority representatives since they cluster with minority examples and the classifiers would suffer to distinguish between them. Then, we employed the randomized undersampling method to further reduce the number of majority examples in order to achieve a balanced distribution.

Finally, we introduced three more strategies to reduce the excessive number of majority examples that comprise the majority class. For each majority example, we added a new attribute that expressed its temporal distance to the closest minority example. Then, we selected only the majority examples that had a temporal distance greater than one day (D1U), or two days (D2U), or four days (D4U). And finally, similarly to strategy CU, we employed in the D1U and D2U strategies the randomized undersampling method to further reduce the number of majority examples and achieve a balanced distribution. In the case of the D4U strategy, the number of the majority examples after the reduction was very close to the number of the minority examples. The idea of the temporal distance arose from the fact that during the precipitation episodes there may be some intervals without precipitation on the ground, while the meteorological factor for the precipitation still exists. It is possible that the classifiers can not distinguish these cases of majority class from a minority one leading to a degradation of their performance.

In Section 5, we provide the corresponding number of examples for each strategy and the details regarding the sub-clusters of the precipitation class variable. The training datasets were the input to five classifiers, namely, the Decision tree C4.5, the k-Nearest Neighbor, the Multi-layer Perceptron with back-propagation, the Naïve Bayesian and the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [21].

We evaluated the resulting models on separate test datasets that followed the natural distribution regarding the clusters of precipitation (Tables 1 and 2). The Area Under the ROC Curve, or simply AUC, was the evaluation metric we used. The AUC measures the performance of the classifiers as a single scalar. ROC graphs are two-dimensional graphs

in which the True Positive Rate (the percentage of minority cases correctly classified as belonging to the minority class) is plotted on the Y axis and the False Positive Rate (the percentage of majority cases misclassified as belonging to the minority class) is plotted on the X axis. An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). The AUC is a reliable measure especially for imbalanced datasets to get a score for the general performance of a classifier and to compare it to that of another classifier [5,16].

5. Experiments and Results

5.1. Training/Test Datasets

The training/test set method was used to build and evaluate the data mining models. The initial dataset of 61,364 examples was divided into 10 non-overlapping folds. By taking each one of the 10 folds as a test set and the remaining 9 as a pool of examples for choosing the training datasets, we formed 10 groups with 55,228 training examples and 6136 test examples. Every fold was chosen randomly, but it followed the natural distribution according to the clusters within the precipitation class variable, as shown in Tables 1 and 2. Thus, we produced 10 test datasets with 6136 examples following the natural distribution that covered the entire initial dataset. In our experiments we always used the above test datasets without introducing any synthetic examples.

We created 100 training datasets by randomly taking 10 samples with replacement consisting of 17,788 examples from the training examples of each one of the 10 groups. Furthermore, we joined the same test dataset 10 times to the corresponding 10 training datasets of each group and formed 100 training/test datasets with 23,924 examples (17,788 training and 6136 test examples, 74.35–25.65%). It is noted that in the strategy of D4U, where we used the four days restriction and reduced the number of majority examples close to the number of minority examples, we formed only a total of 10 training datasets, one for each group.

The different methodologies used to generate a training dataset, characterize the different strategies that we followed to address the class imbalance problem. We employed nine new training datasets according to the strategies that we described in Section 4.

Table 6 shows the number of examples of each of the five different types of precipitation for: (a) the initial file (Initial), (b) the 10 groups (Groups), (c) the 10 folds or test sets (Folds), and, (d) the sampled training datasets produced by the nine strategies. Notice that for all strategies, we generated 10 samples per Group for a total of 100 samples of 17,788 examples. The exception was D4U, where the generated testing datasets had an almost balanced distribution of the majority and minority classes, hence, we generated a single sample per Group for a total of 10 samples of 17,625 examples.

In Table 6, we observe that the total number of minority examples in the original training datasets (Groups of 9 folds) was 8894. Hence, in order to produce a 50% balanced training dataset, one needs to choose the same number of majority examples out of the 46,334 available ones. This is the reason we chose 17,788 as the size of the sampled training dataset. These examples correspond to about 12 years of data that is an acceptable amount of data for classification purposes according to our previous research [3], as we explained in Section 1.

5.2. Algorithm Runs

To recap, we tested each one of the first nine strategies with 100 training/test datasets (UN, U30, U50, SU30, SU50, BW, CU, D1U and D2U) and the tenth strategy with 10 training/test datasets (D4U), for a total of 910 training/test datasets.

These datasets comprised the input to the five classifiers that were run and evaluated using WEKA. The classifiers were the decision tree C4.5 without pruning and Laplace estimate (DT), the k-Nearest Neighbors with k = 5 and Euclidean distance (kNN), the RIPPER (RIP), the Naïve Bayesian (NB), and the Multilayer Perceptron neural network with back-propagation (MP).

Table 6. The natural distribution and the number of examples within the precipitation class variable of the training datasets generated by the various sampling strategies.

	Precipitation 'Yes'				Precipitation 'No'			
	Rain/Drizzle 11.66%	**Snow/Sleet 0.89%**	**Thunder 3.55%**	**Total 'Yes' 16.10%**	**Fog 2.27%**	**Fair/Cloudy 81.62%**	**Total 'No' 83.90%**	**Total 100.00%**
Initial	7154	547	2181	9882	1395	50,087	51,482	61,364
Groups	6438	493	1963	8894	1256	45,078	46,334	55,228
Folds	715	55	218	988	140	5008	5148	6136
Strategy								
UN	2069	171	639	2879	443	14,466	14,909	17,788
U30	3863	296	1177	5336	362	12,090	12,452	17,788
U50	6438	493	1963	8894	250	8644	8894	17,788
SU30	3863	296	1177	5336	362	12,090	12,452	17,788
SU50	6438	493	1963	8894	250	8644	8894	17,788
BW	2965	2964	2965	8894	4447	4447	8894	17,788
CU	6438	493	1963	8894	250	8644	8894	17,788
D1U	6438	493	1963	8894	250	8644	8894	17,788
D2U	6438	493	1963	8894	250	8644	8894	17,788
D4U	6438	493	1963	8894	193	8538	8731	17,625

The last three classifiers were run using the default settings of WEKA. Thus, we performed 4550 runs in the WEKA environment and we present the results in Table 7 and in Figures 1 and 2. Table 7 shows the mean value and the standard deviation of AUC of the 100 or 10 (for D4U) runs for each strategy and classifier.

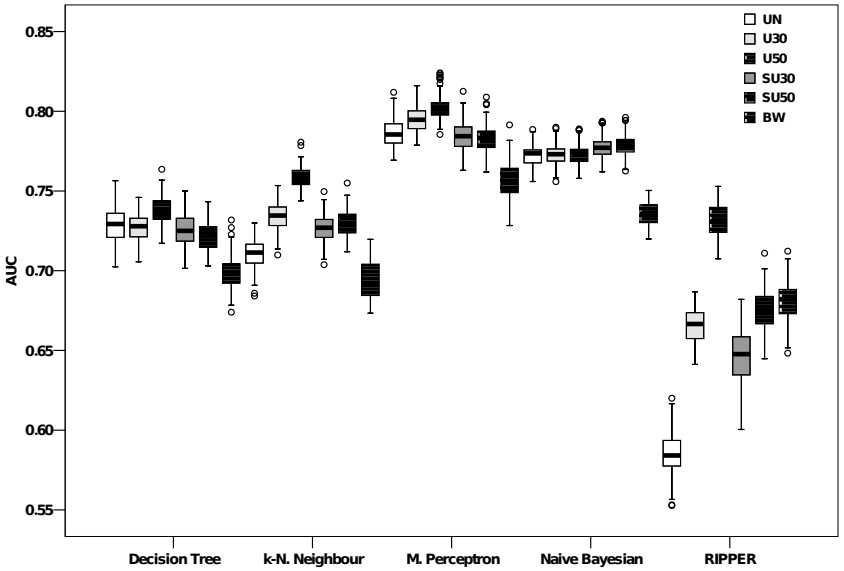

Figure 1. Box-plots of AUC values for strategies UN, U30, U50, SU30, SU50, BW and all classifiers.

Since it is impossible to plot all the box plots for all strategies and classifiers in a single figure, we decided to use two figures. In the first figure, we compare the strategies commonly used in the literature (2 through 6) against UN (strategy 1 that simply uses the initial unbalanced dataset). In the second figure, we compare the newly proposed strategies (7 through 10) against UN and the best strategy of the first figure.

Thus, Figure 1 depicts the box-plots of the corresponding AUC values for the first six strategies. The white box-plots correspond to the UN strategy, the light gray box-plots to the U30 strategy, the light gray box-plots with a pattern of black dots to the U50 strategy, the dark gray box-plots to the SU30 strategy, the dark gray box-plots with a pattern of

black dots to the SU50 strategy and the white box-plots with a pattern of black dots to the BW strategy.

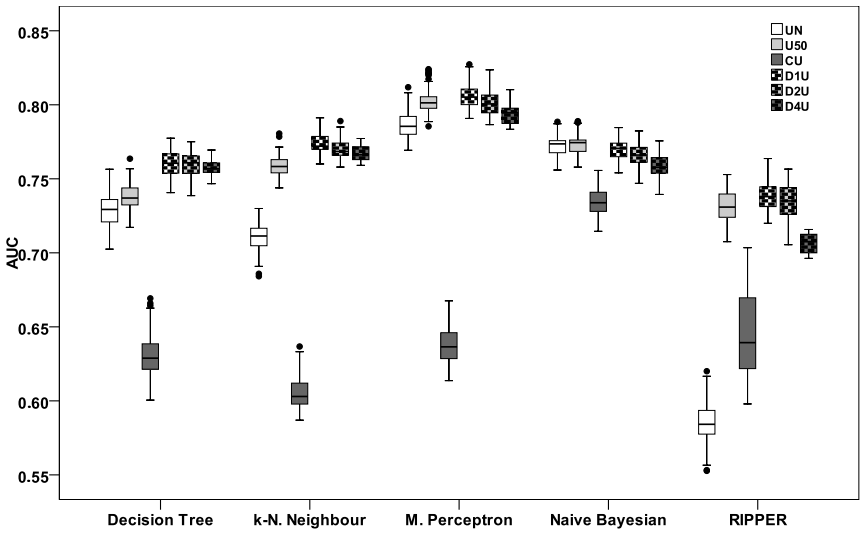

Figure 2. Box-plots of AUC values for strategies UN, U50, CU, D1U, D2U, D4U and all classifiers.

Table 7. Mean value and standard deviation (SD) of AUC. The top three strategies per classifier are shown in red text.

		Classifier				
Strategy		**DT**	**kNN**	**MP**	**NB**	**RIP**
UN	Mean	0.728	0.711	0.786	0.773	0.586
	SD	0.011	0.009	0.009	0.008	0.014
U30	Mean	0.727	0.734	0.795	0.773	0.665
	SD	0.009	0.009	0.009	0.008	0.012
U50	Mean	0.737	0.759	**0.803**	**0.774**	**0.732**
	SD	0.008	0.007	0.008	0.008	0.01
SU30	Mean	0.725	0.726	0.785	**0.778**	0.647
	SD	0.01	0.009	0.009	0.008	0.016
SU50	Mean	0.722	0.73	0.783	**0.779**	0.676
	SD	0.009	0.009	0.009	0.008	0.013
BW	Mean	0.699	0.71	0.757	0.735	0.68
	SD	0.011	0.012	0.012	0.008	0.012
CU	Mean	0.631	0.607	0.638	0.735	0.644
	SD	0.015	0.012	0.013	0.011	0.028
D1U	Mean	**0.76**	**0.774**	**0.806**	0.77	**0.739**
	SD	0.008	0.007	0.008	0.008	0.01
D2U	Mean	**0.759**	**0.77**	**0.802**	0.766	**0.735**
	SD	0.008	0.006	0.008	0.009	0.012
D4U	Mean	**0.757**	**0.768**	0.795	0.759	0.706
	SD	0.007	0.006	0.008	0.009	0.007

We notice that the best strategy for each classifier, with the exception of Naïve Bayesian, is the Randomized Undersampling with the balanced distribution (U50). Also, the classifier with the highest AUC value is the Multilayer Perceptron with back-propagation Neural Network. Regarding the Naïve Bayesian classifier, all strategies perform about equally and it seems that only the combination of the SMOTE Oversampling and Randomized Undersampling strategies (SU30, SU50) slightly improve the AUC metric. For the k-Nearest Neighbor and RIPPER classifiers, the U30, U50, SU30 and SU50 strategies significantly improve the performance on AUC, and, especially, the U50 strategy. For the Decision Tree C4.5, only the U50 strategy performs slightly better than the Natural one (UN), and, for the Multilayer Perceptron, the U50 strategy performs better than the Natural one (UN) and the U30 strategy slightly better. The balanced distribution in both the between and within-classes (BW) strategy gave the worst results on AUC with the exception of the RIPPER classifier.

Likewise, Figure 2, depicts the box-plots of the corresponding AUC values for the proposed four strategies (CU, D1U, D2U, D4U), and, additionally, the UN and U50 strategies for comparison. The U50 strategy was chosen because of its performance shown in Table 7 and Figure 1. The white box-plots correspond to the UN strategy, the light gray box-plots to the U50 strategy, the dark gray box-plots to the CU strategy, the white box-plots with a pattern of black dots to the D1U strategy, the light gray box-plots with a pattern of black dots to the D2U strategy, and the dark gray box-plots with a pattern of black dots to the D4U strategy.

In both Figure 2 and Table 7 that highlights the top three performing strategies per classifier, we notice that the strategies with the temporal distance restriction of each minority example from the closer majority one (D1U, D2U and D4U) perform better than the UN strategy on all classifiers with the exception of the Naïve Bayesian classifier. In addition, they perform better than the U50 strategy in the case of the Decision Tree C4.5 and the k-Nearest Neighbor classifiers. Regarding the Multi-layer Perceptron, Naïve Bayesian and RIPPER classifiers, the D1U strategy performs about equally to or slightly better than the U50 strategy, while it performs better than the D4U strategy. Finally, the CU strategy gave very poor results on AUC and only in the RIPPER classifier it outperformed the UN strategy.

6. Conclusions

We applied Principal Component Analysis to reduce the 71 initial chosen variables of the ERA-40 dataset to 9 variables that were uncorrelated to each other, which explain nearly 85.2% of the variability in the original variables. The reduced ERA-40 dataset and the historical precipitation records of the Meteorological Station of Micra, Thessaloniki, Greece were then input into five data mining and machine learning classifiers we used to build models that predict the occurrence of precipitation at the station.

The Multilayer Perceptron with back-propagation neural network classifier outperforms all other classifiers on AUC, revealing the most effective classifier in this meteorological domain.

Moreover, the proposed new strategy D1U with the balanced distribution resulting from the combination of the one day restriction and the Randomized Undersampling method is the recommended strategy to address the class imbalance problem for the Multilayer Perceptron with back-propagation neural network, Decision Tree C4.5, k-Nearest Neighbor and RIPPER classifiers. Alternatively, the Randomized Under-sampling with the balanced distribution strategy U50 could also be used for the Multilayer Perceptron with back-propagation neural network and RIPPER classifiers. Finally, regarding the Naïve Bayesian classifier, the proposed sampling strategies did not improve its performance when compared to the natural distribution. We observe that in the class imbalance problem, the application of sampling strategies based on the expertise on the data domain can improve the effectiveness of some classifiers.

Author Contributions: Conceptualization, E.T. and G.E.; Methodology, E.T. and G.E.; Software, E.T.; Validation, E.T. and G.E.; Formal analysis, E.T. and G.E.; Investigation, E.T.; Resources, E.T.; Data curation, E.T.; Writing—original draft, E.T.; Writing—review & editing, E.T. and G.E.; Visualization, E.T. and G.E.; Supervision, G.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We wish to thank the European Centre for Medium-Range Weather Forecasts and the Greek National Meteorological Service for providing us with the meteorological data. We would also like to thank our colleagues Demetrios Papanastasiou and Leonidas Karamitopoulos for their valuable suggestions and comments.

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

References

1. Brownlee, J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*; Machine Learning Mastery: San Juan, PR, USA, 2020.
2. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
3. Tsagalidis, E.; Evangelidis, G. The Effect of Training Set Selection in Meteorological Data Mining. In Proceedings of the IEEE 14th Panhellenic Conference on Informatics (PCI 2010), Tripoli, Greece, 10–12 September 2010; pp. 61–65. [CrossRef]
4. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 1–6. [CrossRef]
5. Weiss, G.M. Mining with rarity: A unifying framework. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 7–19. [CrossRef]
6. Healy, D.; Mohammed, Z.; Kanwal, N.; Asghar, M.N.; Ansari, M.S. Deep Learning Model for Thunderstorm Prediction with Class Imbalance Data. In Proceedings of the 15th International Conference on Information Technology and Applications, Dubai, United Arab Emirates, 13–14 November 2021; Ullah, A., Anwar, S., Rocha, Á., Gill, S., Eds.; Springer Nature: Singapore, 2022; pp. 195–205.
7. Bouget, V.; Béréziat, D.; Brajard, J.; Charantonis, A.; Filoche, A. Fusion of Rain Radar Images and Wind Forecasts in a Deep Learning Model Applied to Rain Nowcasting. *Remote Sens.* **2021**, *13*, 246. [CrossRef]
8. Wang, D.; Ding, W.; Yu, K.; Wu, X.; Chen, P.; Small, D.L.; Islam, S. Towards Long-Lead Forecasting of Extreme Flood Events: A Data Mining Framework for Precipitation Cluster Precursors Identification. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13), Chicago, IL, USA, 11–14 August 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 1285–1293. [CrossRef]
9. Nugroho, A.; Kuroyanagi, S.; Iwata, A. Fog forecasting using self growing neural network "CombNET-II"—A solution for imbalanced training sets problem. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 27–27 July 2000; Volume 4, pp. 429–434. [CrossRef]
10. Li, J.; Zhu, Q.; Wu, Q.; Zhang, Z.; Gong, Y.; He, Z.; Zhu, F. SMOTE-NaN-DE: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution. *Knowl. Based Syst.* **2021**, *223*, 107056. [CrossRef]
11. Li, Z.; Qin, J.; Zhang, X.; Wan, Y. Addressing Class Overlap under Imbalanced Distribution: An Improved Method and Two Metrics. *Symmetry* **2021**, *13*, 1649. [CrossRef]
12. Li, Y.; Guo, H.; Zhang, Q.; Gu, M.; Yang, J. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowl. Based Syst.* **2018**, *160*, 1–15. [CrossRef]
13. Hirsch, V.; Reimann, P.; Mitschang, B. Exploiting Domain Knowledge to address Multi-Class Imbalance and a Heterogeneous Feature Space in Classification Tasks for Manufacturing Data. *Proc. VLDB Endow.* **2020**, *13*, 3258–3271. . [CrossRef]
14. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
15. Jo, T.; Japkowicz, N. Class imbalances versus small disjuncts. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 40–49. [CrossRef]
16. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
17. Kållberg, P.; Simmons, A.; Uppala, S.; Fuentes, M. *The ERA-40 Archive*; ECMWF: Reading, UK, 2004; p. 31.
18. Tsagalidis, E.; Evangelidis, G. Pre-processing of Meteorological Data in Knowledge Discovery. In Proceedings of the 10th International Conference of Meteorology, Climatology and Atmospheric Physics, COMECAP 2010, Patras, Greece, 25–28 May 2010.

19. Hair, J.; Black, W.; Babin, B.; Anderson, R. *Multivariate Data Analysis*; Always Learning; Pearson Education Limited: New York, NY, USA, 2013.
20. Hall, M.A.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
21. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann, Elsevier: Amsterdam, The Netherlands, 2011.