The effect of Training Set selection in Meteorological Data Mining

Evangelos Tsagalidis, Georgios Evangelidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
e-mail: {vangelis, gevan}@uom.gr

***Abstract*—We use meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Meteorological Station of Mikra (Thessaloniki, Greece) as input to five data mining algorithms with the aim to build classification models for the prediction of the occurrence of precipitation in the station. We focus our study on the effect the selection of the training set has on the performance of the algorithms and more specifically, we attempt to determine the minimum training set size that can ensure effective application of the data mining techniques.**

Keywords – training set size, meteorological data mining, classification

I. INTRODUCTION

This paper is an application of Knowledge Discovery from Meteorological Databases [1]. We focus our study on the effect of training set selection and size on the efficiency of data mining algorithms in a classification problem. We combine two datasets, the ERA-40 data from the reanalysis project of the European Centre for Medium-Range Weather Forecasts (ECMWF), and the weather observations data from the Meteorological Station of Mikra (Thessaloniki, Greece). Our goal is to predict the occurrence of precipitation in the station – presence or absence of precipitation on the ground – based on a set of ERA-40 parameter values. We apply five different data mining algorithms to construct models for predicting our binary class variable.

In addition, we follow two different strategies for selecting the training datasets. We chose randomly either instances from all years or full years, that is, all the instances of certain years, in order to examine possible differences on the performance of the classifiers.

The paper is organized as follows. Section 2 describes the datasets we used for applying the data mining algorithms. Section 3 discusses the methodology used in the experiments. In Section 4, we present the analysis and the results, and, finally, we conclude in Section 5.

II. DATASETS

The European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis ERA-40 is a global atmospheric analysis of many conventional observations and satellite data streams for the period September 1957 to August 2002. Over the past decade, reanalysis of multi-decadal series of past observations have become an important and widely utilized resource for the study of atmospheric and oceanic processes and predictability [2]. 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. The main objective of the reanalysis project ERA-40 is to promote the use of global analyses of the state of the atmosphere, land and surface conditions over the period 1957-2002 [2].

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.

We created our initial dataset by choosing the values of 10 variables on 7 pressure surfaces on one node. We used only the data from node with geographical coordinates 40^0N latitude and 22.5^0E longitude, which is the closest node to the Meteorological Station of Mikra, Thessaloniki, Greece. The 10 variables are the geopotential in $m^2 \cdot s^{-2}$, temperature in K, U velocity in $m \cdot s^{-1}$, V velocity in $m \cdot s^{-1}$, specific humidity in $kg \cdot kg^{-1}$, relative humidity as percentage (%), vorticity (relative) in s^{-1}, potential vorticity in $K \cdot m^2 \cdot kg^{-1} \cdot s^{-1}$, divergence in s^{-1}, and vertical velocity in $Pa \cdot s^{-1}$. We omit the 11th Ozone mass mixing ratio. The 1000hPa, 925hPa, 850hPa, 700hPa, 500hPa, 300hPa and 200hPa 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 was converted to $g \cdot kg^{-1}$ and vertical velocity to $hPa \cdot h^{-1}$. We multiplied the relatively small values of both vorticity (relative) and divergence by 10^6, and the value of potential vorticity by 10^8. Regarding the wind, wind direction in azimuth degrees and wind speed in knots were calculated

using the U and V velocities. 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 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 and relative humidity on pressure surfaces remained unchanged.

The 6-hourly main synoptic surface observation data of the Mikra Meteorological Station, located at 40.52^0N, 22.97^0E and altitude of 4m, completed our initial dataset. More specifically, we collected the recorded precipitation data of the period 1/1/1960 00UTC – 31/12/2001 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. 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 has priority over rain. Tables I and II depict the distribution of the precipitation types that had been recorded in the Mikra Meteorological Station according to the defined sub-clusters.

III. METHODOLOGY

Firstly, 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 [3], [4].

Then, we examined the effect of the training set size on the efficiency of classification algorithms. It is known that the amount of training data affects the performance of a classifier [5], [6], [7]. Our study focused on the existence of a minimum training set size that could be used to ensure adequate performance of the different mining algorithms.

In addition, we followed two different strategies to build our training datasets. The annual periodicity of the meteorological conditions in a location, prompted us to build training datasets by randomly choosing as groups entire sets of instances belonging to a year; we call this *strategy of Years*. As an alternative strategy, we randomly sampled instances from the complete initial dataset ignoring the temporal dimension of the data; we call this *strategy of Instances.*

The training datasets were the input to five data mining algorithms, 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) [1], [5]. We evaluated the created models on separate test datasets that followed the natural distribution regarding the clusters of precipitation (Tables I and II).

TABLE I. NATURAL DISTRIBUTION OF VALUES WITHIN THE PRECIPITATION CLASS VARIABLE 'YES'

Precipitation 'yes'			
Rain, Drizzle	***Snow, Sleet***	***Thunder***	***Total 'yes'***
7154	547	2181	9882
11.66%	0.89%	3.55%	16.1%

TABLE II. NATURAL DISTRIBUTION OF VALUES WITHIN THE PRECIPITATION CLASS VARIABLE 'NO'

Precipitation 'no'		
Fog	***Fair, Cloudy***	***Total 'no'***
1395	50087	51482
2.27%	81.6%	83.9%

As evaluation metric we used the Area Under the ROC Curve or simply AUC. The AUC measures the performance of the algorithms as a single scalar. ROC graphs are two-dimensional graphs in which the True Positive Rate (the percentage of positive cases correctly classified as belonging to the positive class) is plotted on the Y axis and the False Positive Rate (the percentage of negative cases misclassified as belonging to the positive class) is plotted on the X axis. A ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). The AUC is a reliable measure to get a score for the general performance of a classifier and to compare it to that of another classifier [8], [9], especially for imbalanced datasets like our dataset shown in Tables I and II.

IV. EXPERIMENTS AND RESULTS

A. Feature Selection

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 respecified six times with a final outcome of 35 variables and 9 components with eigenvalues greater than 1. The first 9 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 [3], [4], [10]. The variables that could represent the 9 principal components were the following: geopotential height on 200hPa, relative vorticity on 1000hPa, wind direction on 300hPa, wind speed on 300hPa and on 925hPa, divergence on 300hPa, temperature on 200hPa, potential vorticity on 500hPa and wind direction on 925hPa [4]. 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 61364 examples or instances. We have four instances per day for a time period of 42 years, i.e., 4x365x42 = 61320 instances, plus 44 instances for the 11 leap years.

B. Training – Test datasets

The training/test set method was used to build and evaluate the data mining models. Following the *strategy of Instances*, the initial dataset with 61364 instances was divided into 7 non-overlapping folds. By taking each one of the 7 folds as test set and the rest 6 as a pool of instances for choosing the training sets, we formed 7 groups with 52598 training instances (corresponding to data equivalent to 36 years) and 8766 test instances (corresponding to data equivalent to 6 years). Every fold was chosen randomly, but it followed the natural distribution according to the clusters within the precipitation class variable, as shown in Tables I and II. Thus, we produced 7 test datasets with 8766 instances following the natural distribution that covered the entire initial dataset.

Likewise, following the *strategy of Years*, the initial 42 years dataset (61364 instances) was divided into 7 non-overlapping folds, where each fold consisted of the entire set of instances of 6 years. By taking each one of the 7 folds as test set and the rest 6 folds as a pool of instances (grouped by year) for choosing the training sets, we formed 7 groups with 36 training years (52598 instances) and 6 test years (8766 instances).

Regarding the training data in the *strategy of Instances*, we created 630 datasets (15 x 7 x 6) following the natural distribution according to the percentages shown in Tables I and II. This was accomplished by taking randomly 15 samples with replacement from the training instances of each one of the 7 groups. To build training sets with 6 different sizes of 3, 6, 9, 12, 15, or 18 years we took each time 4386, 8766, 13150, 17538, 21918, or 26302 instances respectively (see Table III). Furthermore, for each one of the 7 groups, we joined 90 times (15 runs x 6 set sizes) the test dataset belonging to the group to the corresponding 90 training datasets of each group.

Regarding the training data in the *strategy of Years*, we created again 630 datasets (15 x 7 x 6) by taking randomly 15 samples with replacement from the training instances, grouped by year, of each one of the 7 groups. This was repeated 6 times, taking 630 (105 x 6) training samples consisting of the entire set of instances of 3, 6, 9, 12, 15, or 18 years (see Table III). Furthermore, we joined 90 times the same test dataset to the corresponding 90 training datasets of each group as in the *strategy of Instances*.

Table III depicts the training and test set sizes and the corresponding percentages we used to build the input to the algorithms for both the *strategies of Instances and Years.*

TABLE III. TRAINING/TEST SETS SIZES FOR THE STRATEGIES OF YEARS AND INSTANCES AND THE CORRESPONDING PERCENTAGES

Training set size			Test set size		
Years	***Instances***	***Pct.***	***Years***	***Instances***	***Pct.***
3	4386	33.3%	6	8766	66.7%
6	8766	50.0%	6	8766	50.0%
9	13150	60.0%	6	8766	40.0%
12	17538	66.7%	6	8766	33.3%
15	21918	71.4%	6	8766	28.6%
18	26302	75.0%	6	8766	25.0%

C. Algorithm runs

To recap, we tested each one of the two strategies and each one of the 6 different training set sizes with 105 training/test datasets, for a total of 1260 training/test datasets. These datasets comprised the input to the five data mining algorithms that were run and evaluated using WEKA [11]. The algorithms were the decision tree C4.5 without pruning and Laplace estimate, the k-Nearest Neighbors with k=5 and Euclidean distance, the RIPPER, the Naïve Bayesian, and the Multilayer Perceptron neural network with back-propagation. The last three algorithms were run using the default settings of WEKA [5], [11]. Thus, we performed 6300 runs in the WEKA environment and we show the results in Fig. 1 and Tables IV and V. Tables IV and V present the mean value (MV) and the standard deviation (SD) of AUC of the 105 runs for each algorithm and training set size (T.Size) for the *strategy of Years* and for the *strategy of Instances* respectively.

Fig. 1 depicts the box-plots of the corresponding AUC values. The white box-plots correspond to the training set size of 3 years, the light gray to the size of 6 years, the dark gray to the size of 9 years, the white with a pattern of cross to the size of 12 years, the light gray with a pattern of cross to the size of 15 years and the dark gray with a pattern of cross to the size of 18 years for both *strategy of Years* (y3, y6, y9, y12, y15, y18) and *strategy of Instances* (i3, i6, i9, i12, i15, i18). It is noted, that for each algorithm, we present the 6 box-plots of the *strategy of Years* with increasing training set sizes, followed by the corresponding box-plots of the *strategy of Instances*.

TABLE IV. MEAN VALUE AND STANDARD DEVIATION OF AUC FOR THE STRATEGY OF YEARS

T.Size		D. Tree	k-NN	MPbp	N. Bayes	RIPPER
3	MV	.704	.690	.745	.770	.576
	SD	.012	.012	.012	.011	.019
6	MV	.721	.702	.769	.772	.577
	SD	.010	.009	.010	.010	.017
9	MV	.727	.708	.780	.773	.579
	SD	.008	.009	.008	.010	.014
12	MV	.733	.711	.787	.773	.582
	SD	.009	.009	.007	.009	.014
15	MV	.739	.714	.792	.773	.585
	SD	.007	.010	.007	.010	.013
18	MV	.743	.716	.793	.773	.584
	SD	.007	.008	.007	.009	.013

TABLE V. MEAN VALUE AND STANDARD DEVIATION OF AUC FOR THE STRATEGY OF INSTANCES

T.Size		D. Tree	k-NN	MPbp	N. Bayes	RIPPER
3	MV	.708	.696	.751	.772	.574
	SD	.011	.009	.012	.007	.017
6	MV	.722	.706	.772	.773	.580
	SD	.007	.007	.008	.007	.014
9	MV	.725	.709	.781	.773	.585
	SD	.008	.007	.007	.006	.013
12	MV	.728	.711	.787	.773	.586
	SD	.007	.007	.006	.006	.013
15	MV	.730	.711	.789	.773	.587
	SD	.009	.010	.006	.006	.011
18	MV	.730	.709	.791	.774	.588
	SD	.007	.008	.005	.007	.012

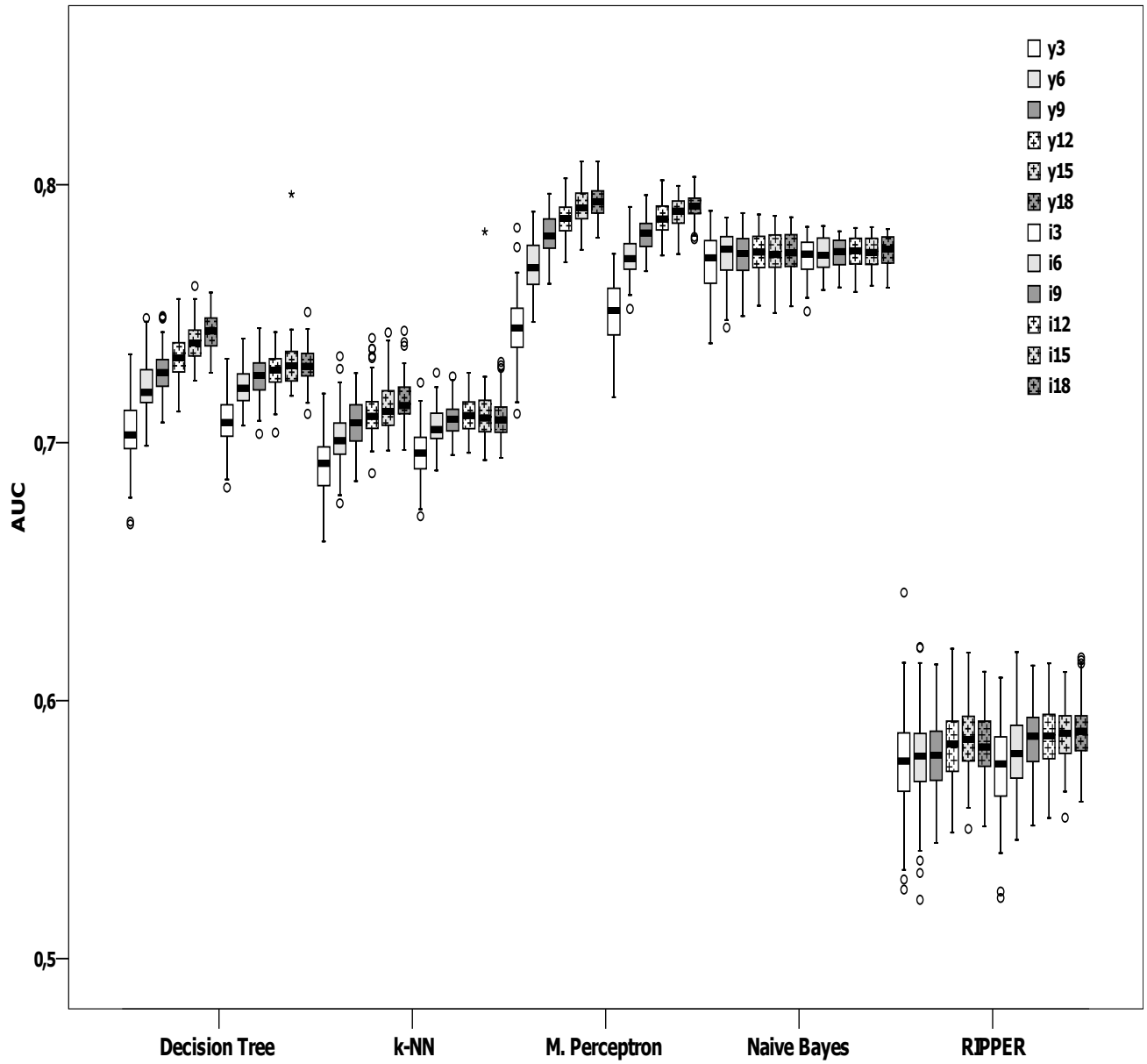

Figure 1. Box-plots of AUC values for each algorithm, strategy and training set size

Comparing the two strategies, we observe that there does not seem to be a significant difference on AUC performance. The decision tree C4.5 algorithm is the only algorithm where the strategy of Years appears to perform slightly better than the strategy of Instances. This is true only for large training set sizes.

Regarding the training set size, for the decision tree C4.5 algorithm and the strategy of Years, the performance on AUC increases gradually with the training set size, whereas, in the strategy of Instances reaches a plateau after the training set size of 9 years. The performance of the k-Nearest Neighbor algorithm on AUC increases slightly with the training set size, but after the size of 9 years reaches also a plateau, especially for the strategy of Instances.

The performance of the Multilayer Perceptron with back-propagation algorithm on AUC rises with the training set size up to the size of 12 years. After that size there are not significant differences. As concern both the Naïve Bayesian and RIPPER algorithms, the performance on AUC remains almost unchanged with the different training set sizes.

The Multilayer Perceptron with back-propagation neural network algorithm outperforms all other algorithms for the relatively large training set sizes (more than 6 years), whereas, the Naïve Bayesian algorithm is better for the relatively small training set sizes (less than 6 years).

V. CONCLUSIONS

The performance on AUC of the decision tree C4.5, k-Nearest Neighbor, and Multilayer Perceptron with back-propagation neural network algorithms does not increase significantly for training set sizes more than 9 years. The performance on AUC of the Naïve Bayesian and RIPPER algorithms is independent from the training set size.

Moreover, the results appear not to be affected by the way we choose the training set, that is, whether we choose randomly isolated instances or the entire set of instances of random years to build the training sets. In this meteorological domain, for training sets with size of more than 6 years the preferable algorithm is the Multilayer Perceptron with back-propagation neural network. For training sets with size of less than 6 years the Naïve Bayesian algorithm appears to be the best choice.

ACKNOWLEDGMENT

We wish to thank the European Centre for Medium-Range Weather Forecasts and the Hellenic National Meteorological Service for providing us with the meteorological data. We would also like to thank Prof Demetrios Papanastassiou and Dr Leonidas Karamitopoulos for their valuable suggestions and comments.

REFERENCES

[1] R. Roiger, M. Geatz, Data Mining: A Tutorial-based Primer. Addison Wesley, USA, 2003.

[2] European Centre for Medium-Range Weather Forecasts (ECMWF, ERA-40). Available: http://www. ecmwf.int/research/era/do/get/era-40

[3] SPSS, Statistical Analysis Software. Available: http://www.spss.com

[4] E. Tsagalidis, G. Evangelidis, "Pre-processing of Meteorological Data in Knowledge Discovery," Proc. 10th International Conference of Meteorology, Climatology and Atmospheric Physics, (Comcap 2010), Patras, Greece, 2010, pp. 61-69.

[5] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques 2nd Ed., San Francisco CA: Morgan Kaufmann, 2005.

[6] T. Jo, N. Japkowicz, "Class Imbalances versus Small Disjuncts," SIGKDD Explorations, vol. 6, issue 1, pp. 40-49, 2004.

[7] G. M. Weiss, "The effect of small disjuncts and class distribution on decision tree learning," Ph.D. dissertation, Graduate School-New Brunswick Rutgers, The State University of New Jersey, 2003.

[8] G. Batista, R. Prati, MC Monard, "A study of the behaviour of several methods for balancing Machine Learning Training Data," SIGKDD Explorations, vol. 6, issue 1, pp. 20-29, 2004.

[9] G. M. Weiss, "Mining with Rarity: A Unifying Framework," SIGKDD Explorations, vol. 6, issue 1, pp. 7-19, 2004.

[10] J. Hair, W. Black, B. Babin, R. Anderson, R. Tatham, Multivariate Data Analysis, 6th Ed., New Jersey: Pearson Prentice Hall, 2006.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, issue 1, 2009.