LEARNING TO PREDICT FOREST FIRES WITH DIFFERENT DATA MINING TECHNIQUES

Daniela Stojanova[1], Panče Panov[2], Andrej Kobler[1], Sašo Džeroski[2], Katerina Taškova[3]
[1]Slovenian Forestry Institute, Ljubljana, Slovenia
[2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[3]Faculty of Information Technology, European University, Skopje, Macedonia
E-mails: stojanovad@yahoo.com, pance.panov@ijs.si,
andrej.kobler@gozdis.si, saso.dzeroski@ijs.si, tkejt@yahoo.com

ABSTRACT

The motivation for this study was to learn to predict forest fires in Slovenia using different data mining techniques. We used predictive models based on data from a GIS (geographical information system), the weather prediction model - Aladin and MODIS satellite data. We examined three different datasets: one only for the Kras region, one for whole Primorska region and one for continental Slovenia. On these datasets we applied logistic regression and decision trees, as well as random forests, bagging and boosting of decision trees, in order to obtain predictive models of fire outbrakes. Best results in terms of predictive accuracy were obtained by bagging decision trees.

1 INTRODUCTION

Forest fires cause significant material damage in the natural environment. A large number of fires is caused by humans, although other factors like drought, wind, topography, plants etc., also have an important indirect influence on fire appearance and its spreading. The fire threat in Slovenia is increasing because of the processes of abandonment of farmland and spontaneous afforestation (increasing the fire fuel) and the increase of recreation in the natural environment.

Fire prevention is the first step in reducing the damage caused by fire and estimation of fire movement is very important for successful fire prevention, organization of prevention measures and optimal storage of firefighting resources. An important tool for fire movement estimation is modeling of the relations between the fire threat and the influence factors. Because these factors are more or less geograpicaly determined, these types of models are usually developed within GIS (Geographical Information System). GIS is a computer system capable of capturing, storing, analyzing, and displaying geographically referenced information; that is, data identified according to their location.

There are already two systems operating in Slovenia used for fire threat assesement in the natural environment. One is operated by the Forest Institute of Slovenia and the other by Slovenian environment agency. The main problem of these systems is that they are outdated and unreliable and tend to spatialy and temporaly over-generalize the results. It is possible to improve the models by enhancing the layers of the input data for the influence factors (GIS part), by increasing the thematic and spatial details, as well as including a database of past fires. The performance of these systems could be improved with employement of better data modeling techniques based on advanced machine learning methods.

The intention of this study is to improve the existing models by including additional data from GIS, ALADIN (**A**ire **L**imitée **A**daptation dynamique **D**éveloppement **I**nterNational), MODIS (Moderate-resolution Imaging Spectroradiometer) and the data on the forest stand height and canopy cover [12] . In addition, we extended the validity of the models to the whole territory of Slovenia.

The rest of the paper is structured as follows. In section 2 we explain the data used in this study. Section 3 describes the data mining techniques used to build the predictive models of fire outbreaks. In section 4 we explain the experimental setup and in section 5 we present and discuss the results. Section 6 gives the conclusions and future work.

2 DESCRIPTION OF THE DATA

In this study we apply data mining techniques and compare their performance (accuracy, precision, recall) to three datasets that contain data from different regions of Slovenia: Kras region in western Slovenia, whole Primorska region, and the continental Slovenia. The task is to predict the fire outbrakes.

The descriptive data is divided into 3 groups:

- GIS data (geographic data, part of the land with forest, field, urban part, distance from roads, highways, railways, cities etc.)
- Multi-temporal MODIS data: daily records for average temperature for specific quadrant and average net primary production. The data covers one year period.
- Meteorological ALADIN data (temperature, humidity, sum energy, evaporation, speed, direction and course of the wind, transpiration etc.)

The spatial measure used was 1x1 km^2 quadrant, and every spatial and time attribute was adjusted to this resolution.

The GIS data contains time independent attributes for every quadrant. The attributes describe the following GIS properties: ID for every quadrant in Slovenia, median of altitude above sea level, median of the grade of relief in gradients referred to DMR100 (Digital Relief Model), modus of exposition of the relief referred to DMR100, distance of roads, distance of cities, distance of railways (if the values are above 15.000 m, 15.000 m is assumed), share of specific land usage in a quadrant (e.g. fields, gardens, forests, buildings and others).

From NASA archives (MODIS satellite) we obtained public data of land temperature and net primary production of plants for the period of 5 years (2000 - 2004) with spatial resolution of 1 km and time resolution of 8 days. Multi-temporal satellite MODIS data, implicitly give information about the response of the vegetation in periods of drought and the types of fire fuels. The data is daily dependant. The MODIS attributes describe the following properties: average temperature in Kelvin for a specific quadrant for the day x of the year (we have 46 values and temperatures for every 8 days. x takes the values of 1, 9, 17, 25, ..., 361), average net primary production for a specific quadrant for the day x of the year.

The ALADIN data contains meteorological predictions of the weather. They are issued daily from the Environmental Agency of the Republic of Slovenia. The data includes weather predictions for every 3 hours (00.00-21-00 UTC) of 10 weather attributes. The attributes are: atmospheric precipitation, sun radiation energy, velocity and direction and gust of wind, evapotranspiration, transpiration, evaporation, relative humidity and temperature. For the three hours of weather data we decided to study the time interval from 12.30 to 15.30. This particular time interval was selected because of the great danger of fire at these hours, as shown from the Aladin data, given in UTC time (winter/summer season is not a factor). An average daily prediction is added to help in removing the noisy data. All of the 10 parameters were averaged for 1, 2, 4 and 14 days.

For the Kras region we also used 3D vegetation data from LIDAR and LANDSAT images. This data contains attributes that describe the height and structure of vegetation in the Kras region. The values of the attributes were calculated from the LANDSAT and were calibrated with LIDAR. All the data are aggregated to 1 km quadrants and have a resolution of 25 x 25 m. The LANDSAT based data describe the following forest properties: average tree canopy height, forest canopy cover, maximum height of the vegetation above the ground, height above the ground that reaches some percentage (99%, 95%, 75%, 50%) of forest biomass. For every attribute we obtain 4 statistic measures – minimum, maximum, average and standard deviation on 1km quadrant.

For building predictive models of forest fires we need positive and negative examples of fire occurrences. Positive examples of fires are locations in the past, where we have noticed the fire occurrence along with the date and hour. Negative examples are represented by an equal number of points with random time stamps and randomly located within the areas at least 15 km away from any positive example detected in timestamp ± 3 days [11]. This algorithm gives precedence to the area that had smaller probability of fire occurrence in a defined period. Locations of the positive and negative examples of fire occurrence were spatially and temporally linked to the descriptive data. The data for locations of fire occurrences (positive examples) were obtained from the Administration for Civil Protection and Disaster Relief of Slovenia and the Forestry Institute.

3 DATA ANALYSIS METHODOLOGY

The data were analyzed with several different data mining algorithms for classification implemented in WEKA data mining system [4].We used: logistic

regression, random forests, decision trees (J48), bagging and boosting ensemble methods.

Logistic regression is part of a category of statistical models called generalized linear models [6]. Logistic regression allows prediction of a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. There are two main uses of logistic regression. The first is the prediction of group membership. Since logistic regression calculates the probability or success over the probability of failure, the results of the analysis are in the form of an odds ratio. Logistic regression also provides knowledge of the relationships and strengths among the variables.

Random forest [10] is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier.

J48 algorithm [4] is an implementation of the C4.5 decision tree learner [5]. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. There are many parameters which can be tuned in order to obtain better models with respect to the accuracy (or other parameters which can be used as measure for the quality of the model). These parameters allow greater control of the user in the process of learning the models.

Often multiple versions of a classifier give better results than the individual base classifier, because of combining the advantages of the individual classifiers in the final (aggregated) classifier. The simplest way to do the "aggregation" in the case of classification is to take a vote (perhaps a weighted vote); in the case of numeric prediction, to calculate the average (perhaps a weighted average).

Bagging predictors [7] is a method for generating multiple versions of a predictor (making bootstrap replicates of the learning set and using these as new learning set) and using them to get an aggregated predictor. In bagging all models receive equal weight. Bagging produces very accurate probability estimates from decision trees and other powerful, yet unstable, classifiers. However, a disadvantage is that bagged classifiers are hard to interpret.

Boosting is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule [8]. The boosting algorithm calls this "weak" or "base" learning algorithm repeatedly, each time feeding it a different subset of the training examples (or, to be more precise, a different distribution or weighting over the training examples). A widely used method for boosting is AdaBoost [9]. AdaBoost calls a given weak or base learning algorithm repeatedly in a series of rounds. The algorithm maintains a distribution or set of weights over the training set. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the base learner is forced to focus on the hard examples in the training set. At the end predictions of all weak rules are combined into a single prediction with weighted voting.

4 EXPERIMENTAL SETUP

As it was described in Section 2 the purpose of this study is to learn predictive models of forest fires outbreaks. The experiments are performed on three datasets for different regions of Slovenia: Kras, Primorska and continental Slovenia. The Kras dataset contains 159 attributes and has 1439 examples. The Primorska dataset has 129 attributes and 2442 examples. The third dataset for continental Slovenia has 129 attributes and 8476 examples. For all datasets the target attribute is nominal and predicts the possibility of fire occurrence (0-no, 1- yes). For conducing the experiments we used WEKA [4] data mining system. Several algorithms were used in the experiments: logistic regression, random forests, J48 (WEKA's implementation of decision trees), bagging and boosting of trees.

All of the methods were used with the default parameters. Ensemble methods were run in 10 iterations. The validation of the models was done using 10 fold cross – validation.

5 RESULTS AND DISCUSION

In this section, we present the results we obtained from the experiments. In Tables 1-3 we present the performances of the experiments in terms of precision, recall, accuracy and kappa statistics for each of the datasets respectively. Precision and recall in this case are calculated for the class 1 (fire

occurrence=yes). Kappa statistics is used to evaluate the agreement between predicted and observed nominal values in one dataset, while correcting for agreement that occurs by chance.

Algorithm	Precision	Recall	Accuracy	Kappa
Logistic reg.	0.696	0.563	0.772	0.461
Random F.	0.751	0.585	0.797	0.517
J48	0.639	0.652	0.761	0.465
Bagging	**0.754**	0.652	**0.812**	**0.560**
Boosting	0.725	**0.658**	0.790	0.520

Table 1 Performances of DM algorithms on Kras data

Algorithm	Precision	Recall	Accuracy	Kappa
Logistic reg.	0.826	0.849	0.834	0.668
Random F.	0.820	**0.903**	0.852	0.703
J48	0.810	0.810	0.809	0.618
Bagging	**0.850**	0.878	**0.860**	**0.721**
Boosting	0.839	0.867	0.856	0.712

Table 2 Performances of DM algorithms on Primorska data

Algorithm	Precision	Recall	Accuracy	Kappa
Logistic reg.	0.831	0.855	0.840	0.679
Random F.	0.823	**0.877**	0.843	0.703
J48	0.809	0.819	0.812	0.624
Bagging	**0.846**	0.856	**0.849**	**0.698**
Boosting	0.842	0.855	0.844	0.688

Table 3 Performances of DM algorithms on continental Slovenia data

From the results we can conclude that Bagging of decision trees shows the best results in terms of predictive accuracy, precision and kappa statistics compared to the other algorithms.

5 CONCLUSION AND FUTURE WORK

In this work we built predictive models of forest fires. The models were based on GIS, MODIS and Aladin data. The experimental results showed that bagging of decision trees gives the best results in terms of accuracy for all three datasets. In further work we would like to use some feature selection algorithms and try to extract the relevant features and try to further improve the accuracy of the models.

References

[1] SFS Slovenian Forestry Service. Slovenian forest cover statistics 2004, unpublished manuscript. Zavod za gozdove RS, Ljubljana, Slovenia, 2006.

[2] FAO-Food and Agriculture Organization of the United Nations. Global Forest Resources Assessment Update 2005, Slovenia Country Report. 2005.

[3] R.M. Measures. Laser remote sensing: fundamentals and applications. Malabar, Fla., Krieger Pub. Co., 1992. 510 p. G70.6M4, 1992.

[4] I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005. 2nd Edition.

[5] J. R. Quinlan. C4.5: Programs for machine learning. Morgan Kaufmann, 1993.

[6] A. Agresti. An Introduction to Categorical Data Analysis. John Wiley and Sons, Inc., 1996.

[7] L. Breiman. Bagging Predictors. Machine Learning, 26, 123-140, 1996.

[8] R. E. Schapire. The Boosting Approach to Machine Learning – An Overview. MSRI Workshop on Nonlinear Estimation and Classification, 2002.

[9] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55, 119–139, 1997.

[10] L. Breiman. Random Forests. Machine Learning, 45, 5-32, 2001.

[11] A. Kobler, P. Ogrinc, I. Skok, D. Fajfar, S. Džeroski. Končno poročilo o rezultatih raziskovalnega projekta: Napovedovalni GIS model požarne ogroženosti naravnega okolja

[12] S. Džeroski, A. Kobler, V. Gjorgijoski, P. Panov. Using Decision Trees to Predict Forest Stand Heght and Canopy Cover from LANDSAT and LIDAR data. 20th International Conference on Informatics for Environmental Protection - Managing Environmental Knowledge – ENVIROINFO 200