Hail Size Estimation and Prediction using Data Mining Techniques

Evangelos G. Tsagalidis[1], Kyriakos G. Tsitouridis[1], Georgios Evangelidis[2], and Dimitris A. Dervos[3],
(1) Hellenic Agricultural Insurance Organization, Meteorological Applications Centre,
International Airport Makedonia, 55103 Thessaloniki, GREECE,
{e.tsagalidis, k.tsitouridis}@elga.gr
2) Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, 54006, Thessaloniki, GREECE, gevan@uom.gr
3) Information Technology Dept., T.E.I.,
P.O. BOX 141, 57400 Sindos, GREECE, dad@it.teithe.gr

Abstract

In this study we examine the existence of interesting patterns among the Greek National Hail Suppression Program data using Data Mining techniques. More specifically, we focus on hail size estimation and prediction from meteorological radar and sounding data. The sought objective is to examine existing relationships and, by doing so, construct a hail size prediction model. Two data mining techniques are applied in order to identify the optimum number of independent variables and, consequently, build a simple, yet effective, model. A model easily applied by the meteorologist in order to quickly interpret radar and atmospheric measurements to possible hail size on the ground.

1. Introduction

The Hellenic Agricultural Insurance Organization (EL.G.A.) is a public organization and the main insurance carrier of the agricultural production in Greece. The Meteorological Applications Centre (KE.M.E.) is the section of EL.G.A which has conducted, since 1981, the Greek National Hail Suppression Program (GNHSP) using airborne seeding, aimed at reducing insurance payments due to hail damage. The Program is being applied in Central Macedonia and Thessaly in the period April to September, covering an area of 5,000 square kilometers. The cloud seeding is performed by three aircraft releasing AgI in developing hail-bearing clouds as indicated by radar [1].

In this study we examine the existence of interesting patterns among the GNHSP data using Data Mining techniques. More specifically, we focus on hail size estimation and prediction from meteorological radar and/or sounding data. Two data mining techniques are applied in order to identify the optimum number of independent variables and, consequently, build a simple, yet effective, model. A model to be used by the meteorologist in order to quickly interpret radar and atmospheric measurements to predict possible hail size on the ground.

The paper is organized as follows: Section 2 describes the dataset used for applying the data mining algorithms that we describe in Section 3. In Section 4 we present the results we obtained by experimenting with the chosen data mining algorithms, and, finally, we conclude the paper in Section 5.

2. Dataset used for Data Mining

The analysis utilizes radar data of the EEC S-band meteorological radar installed at Macedonia Airport of Thessaloniki. The data recorded by the Thunderstorm Identification, Tracking, Analysis and Nowcasting system (TITAN) [2], are further analyzed to create a sample of Storm Cell Complexes (SCC) which is actually a structured form of the initial data and represents the storm characteristics data [3],[4]. The data were recorded during the storm activity from April to September 1999, 2000, 2001 and 2005 in the protected area of Central Macedonia. The SCC structured data represent hailstorm attribute values, more specifically: the *Type*, *Reflectivity*, *Cloud top*, *Vertically Integrated Liquid Water content (VIL)*, *Vertically Integrated Liquid Water Density (VIL Density)* and *Month*.

The structure of cloud systems and their classification in different categories follows the classification of SCC [3],[4]. The classes are represented by the values of the *Type* attribute, where "S" is used for the Unicellular storms of a Single ordinary cell, "SU" for the Unicellular storms of a Supercell, "M" for the Multicell storms, and, "L" for the Line storms. During the entire lifetime of the SCC, the *Reflectivity* in dBz is the maximum radar reflectivity at the -5^0C level or higher and the *Cloud top* in km is the maximum height. The *VIL* in kg·m^{-2} is a function of reflectivity, and converts SCC reflectivity data into an equivalent liquid water content value, whereas the *VIL Density* is simply the VIL divided by the echo top (m) and multiplied by 1000 in order to express the result as g·m^{-3}. Also, the month of occurrence of each SCC event is represented by *Month* attribute. Table 1 lists the mean, standard deviation, minimum and maximum values of *Reflectivity*, *Cloud top*, *VIL* and *VIL Density*.

Table 1: Mean, standard deviation, minimum and maximum values of *Reflectivity, Cloud top, VIL* and *VIL Density*.

	Refl. (dBz)	C. top (km)	VIL (kg·m^{-2})	VIL Density (g·m^{-3})
Mean	52.2	10.4	22	2.2
St.dev.	5.3	1.4	12.1	1.1
Min.	40	7	3.8	0.5
Max.	69	14.5	55.8	5.1

Furthermore, the analysis conducted considered meteorological parameters from the sounding data of the Upper Air Observation Station of Thessaloniki, which relate to the hail size on the ground too. The meteorological station is located close to the project's area of Central Macedonia and the calculated values of the atmospheric parameters, such as Wet Bulb Zero (*WBZ*) and *Mean temperature* are associated with the SCC environment. The *WBZ* is the height in km of the wet bulb temperature 0^0C level, corresponding to the melting level of the hailstone during its fall to the ground, whereas the mean temperature in Kelvin of the layer between that level and the ground is the *Mean temperature*. These parameters were calculated using the most representative sounding as concern the appearance time of each SCC. In the Table 2 are shown the mean, the standard deviation, the minimum and maximum values of *WBZ* and the *Mean temperature* values.

Table 2: Mean, standard deviation, minimum and maximum values of *WBZ* and *Mean temperature*.

	WBZ (m)	Mean temp. (K)
Mean	3123	288
St.dev.	419	2.1
Min.	2118	282
Max.	4182	293

The *WBZ* values associated with hail days are bounded in a range of possible values, because low values of *WBZ* indicate stable air conditions not sufficient for hailstorms, and high values imply an increased possibility to melt the hailstones before reaching the ground [5]. During the preprocessing phase, we made the appropriate transformations of the numerical *WBZ* values to ordinal *'low'*, *'middle'* and *'high'* values using the Z-score normalization method. The *'middle'* value corresponds to a *WBZ* Z-score between -1 and 1, the *'low'* to less than -1 and the *'high'* to greater than 1.

Each SCC is identified as a hailstorm using the data from the GNHSP hailpad network. These data include values of maximum hail diameter in mm produced by each one hailstorm (SCC) as well as the corresponding hail size classes, called *Hail*, such as pea, grape and walnut or more. More specifically, the *Hail* pea class corresponded to a maximum hail diameter recorded in a hailpad between 6 and 12mm, the grape to a maximum hail diameter between 13 and 20mm and the walnut or more to a maximum hail diameter greater than 20mm. Our sample does not include any case whereby two or more SCC passed over the hailpad with the maximum recorded hail diameter.

For each one SCC, the values of the parameters representing one group of radar, sounding, and hailpad network data, were taken to comprise one record. Consequently, 74 records in total were constructed for the 74 SCC identified on radar and having hail records on the hailpad network.

3. Methodology

In the introduction we referred to the aim of the study being the estimation and prediction of hail size, that is associated with a SCC, using radar or/and sounding parameters.

The problem of prediction the hail size or hail class from our database is a typical classification problem. Formally, the classification problem is stated as follows:

Given a database D = {t1, t2, …, tn} of tuples and a set of classes C = {C1, C2, …, Cn}, define a mapping
f: D → C where each ti is assigned to one Ci.

In our case we have tuples consisting of the variables *Month, Reflectivity, Cloud top, Kind, VIL, VIL Density, WBZ, Mean temperature* and the values of the observed *Hail* corresponds to the classes.

3.1 Classification techniques

There is a wealth of classification techniques mainly developed from the fields of Statistics and Machine Learning, such as regression, Bayesian classification, decision trees, nearest neighbor, neural networks, and support vector machines. We chose supervised classification techniques, such as the Decision Tree-based algorithm of C4.5 and the Bayes classifier to perform classification in our dataset.

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Because of their tree structure and ability to easily generate rules decision trees are the favored technique for building understandable models. An algorithm for building decision trees is C4.5 that uses a measure taken from the information theory to help with the attributes selection and data splitting process.

Bayer classifier offers a simple yet powerful supervised classification technique that is based on Bayes theorem:

$$P(H \mid E) = \frac{P(E \mid H) x P(H)}{P(E)}$$

where H is a hypothesis to be tested and E is the evidence associated with the hypothesis. From a classification viewpoint, the hypothesis is the dependent variable and represents the predicted class. The evidence is determined by values of the input attributes. P(E|H) is the conditional probability that H is true given evidence E. P(H) is an *a priori* probability, which denotes the probability of the hypothesis before the presentation of any evidence. Conditional and *a priori* probabilities are easily computed from the training data. The model assumes all input attributes to be of equal importance and independent of one another. Even though these assumptions are likely to be false, the Bayes classifier still works quite well in practice [6].

4. Analysis and results

The WEKA software [7] was used to apply C4.5 (J48) [8] and the Naïve Bayes classifier on our dataset. The two techniques are applied in two modes, (a) validating the method by using 66% of the sample as the training set and the rest of it as the testing sample, (b) using the statistical practice of Cross-validation, partitioning a sample of data

Evangelos Tsagalidis, Hellenic Agricultural Insurance Organization, e.tsagalidis@elga.gr

into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. We experimented with various subsets of the input variables to achieve the best result concerning the prediction of the *Hail* classes. For both techniques best results were obtained by applying the (a) mode of percentage split. Table 3 lists the percentage of correctly classified instances using the C4.5 and Naïve Bayes classification algorithms when applying the (a) mode.

Table 3: Percentage of correctly classified instances for various input variables subsets.

	Correctly classified instances (%)	
Input variables	C4.5	Naïve Bayes
Cloud top, Reflectivity	0.73	0.846
Cloud top, Reflectivity, WBZ, Mean temperature	0.808	0.808
Cloud top, VIL Density	0.692	0.692

We note that the *Cloud top* and *Reflectivity* give the best result using the Naïve Bayes algorithm (84.6%) and the *Cloud top*, *Reflectivity*, *WBZ* and *Mean temperature* using the C4.5 algorithm (80.8%) concerning the percentage of correctly classified instances. Tables 4 and 5 present the corresponding confusion matrix, Precision and Recall values of the Naïve Bayes and the C4.5 algorithms respectively. The confusion matrix shows how many instances of each class have been assigned to each class. Recall expresses how many instances of a class (percentage) have been correctly assigned to that class. Precision expresses how many from the instances assigned to a class (percentage) have been correctly assigned to that class.

Table 4: Confusion matrix of the Naïve Bayes algorithm with *Cloud top* and *Reflectivity* as input variables.

Hail classes	Classified as			Recall
	pea	grape	walnut	
pea	18	0	0	1
grape	3	1	1	0.2
walnut	0	0	3	1
Precision	0.857	1	0.75	

Table 5: Confusion matrix of the C4.5 algorithm with *Cloud top, Reflectivity, WBZ* and *Mean temperature* as input variables.

Hail classes	Classified as			Recall
	pea	grape	walnut	
pea	18	0	0	1
grape	5	0	0	0
walnut	0	0	3	1
Precision	0.783	0	1	

Figure 1 sketches the visualization output of the C4.5 decision tree with *Cloud top*, *Reflectivity*, *WBZ* and *Mean temperature* as input variables and *Hail* as output.

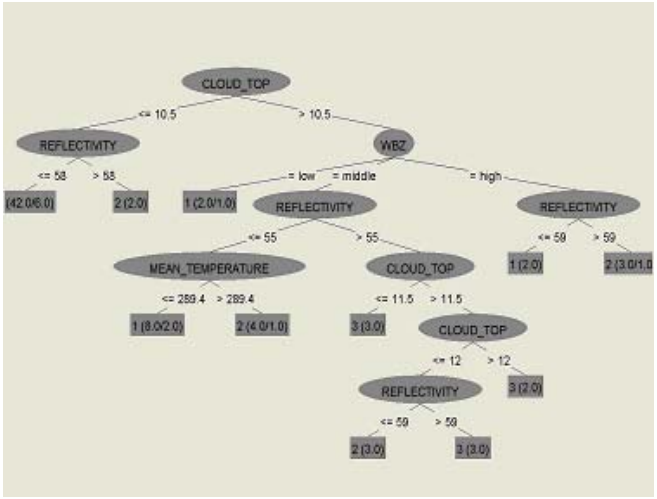

Fig. 1 C4.5 decision tree.

As we remark in all cases above, the efficiency of the models to predict the second *Hail* class *'grape'* is very low, whereas it is high for the other two classes, the *'pea'* and *'walnut'*.

5. Conclusion

The present study is an application of Knowledge Discovery from Databases (KDD) methodology to the specific system of GNHSP. We examine the prediction of hail size, that is associated with a SCC (hailstorm), using radar or/and sounding parameters. The classification techniques of decision tree (C4.5) and Naïve Bayes are used for data mining. The correctly classified instances using the Naïve Bayes algorithm with input variables the *Cloud top* and *Reflectivity* are 84.6% for *Hail* prediction. The first *'pea'* and the third *'walnut'* class are correctly classified, whereas the approach is not efficient in the classification of the middle 'grape' class. In the case of the latter, only one out of five cases have been classified correctly (i.e. a 20% success ratio). Analogous results have been obtained by applying the C4.5 algorithm with *Cloud top, Reflectivity, WBZ* and *Mean temperature* as input variables. The correctly classified instances reached a percentage of 80.8% in correctly classifying the *'pea'* and the *'walnut'* classes, whereas no middle (*'grape'*) class was classified correctly. In the future, we plan to continue experimenting in the direction of improving our model by considering more input variables. More specifically, variables leading to an acceptable classification rate for the middle class (*'grape'*) of hail sizes.

References

[1] S. Tzoumaki, E. Tsagalidis, E. Chatzi, S. Dimoutsi, "Seeding operations in the Greek national hail suppression program", J. Wea. Mod., vol. 38, pp. 16-22, 2006.

[2] M. Dixon, G. Wiener, "TITAN: thunderstorm identification, tracking, analysis, and nowcasting – a radar based methodology", J. Atmos. Ocean. Technol., vol. 10, no. 6 pp. 785-797, 1993.

[3] E. Tsagalidis, K. Tsitouridis, "Storm Cell Complexes: Types, features comparison and regions of appearance", 5th Hellenic Conference of Meteorology, Climatology and Atmospheric Physics, Thessaloniki., September 2000 (Presentation).

[4] E. Tsagalidis, E. Chatzi, D. Boucouvala, "Comparison of the hailstorm characteristics between two different areas in Greece", J. Wea. Mod., vol. 38, pp. 11-15, 2006.

Evangelos Tsagalidis, Hellenic Agricultural Insurance Organization, e.tsagalidis@elga.gr

[5] E. Tsagalidis, “The height of the wet bulb zero temperature (WBZ) and its relationship with hail on the ground during the period April-September 1993”, 3th Hellenic Conference of Meteorology, Climatology and Atmospheric Physics, Athens, September 1996, pp. 117-122, (Presentation).
[6] R. J. Roiger, M. W. Geatz, “Data Mining: A tutorial based primer”, Pearson Education, Inc., 2003.
[7] I. H. Witten, E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, 2nd Edition, Morgan Kaufmann Publishers, June 2005.
[8] J. R. Quinlan, ‘’C4.5: Programs for Machine Learning’’, Morgan Kaufmann Publishers. San Mateo, CA, 1993.

Evangelos Tsagalidis, Hellenic Agricultural Insurance Organization, e.tsagalidis@elga.gr