Prediction of Hail Suppression Program Seeding Parameters using Data Mining Techniques

Evangelos Tsagalidis[1], Georgios Evangelidis[2]

[1] Hellenic Agricultural Insurance Organization, Meteorological Applications Centre, International Airport Macedonia, GR-55103 Thessaloniki, GREECE
Tel: +30 2310472953, Fax: +30 2310472205, E-mail: e.tsagalidis @ elga.gr, vangelis@uom.gr

[2] Dept. of Applied Informatics, University of Macedonia, 156 Egnatia St., GR-54006, Thessaloniki, GREECE, Tel: +30 2310891844, Fax: +30 2310891808, gevan@uom.gr

Abstract

In this study we examine the existence of interesting patterns among the Greek National Hail Suppression Program (GNHSP) data using Data Mining techniques. Two groups of GNHSP data are used. The hailstorms data, containing the values of some hailstorm attributes, such as type, life time, intensity, size, motion and the seeding data, containing the values of some seeding parameters, such as seeding time duration, seeding material mass consumption and mean seeding rate. The results we obtain in the form of association rules can contribute to the prediction of seeding parameters from storm data and the determination of hailstorm characteristics from seeding data.

1. Introduction

In this paper we present an application of the Knowledge Discovery in Databases (KDD) process that attempts to extract useful knowledge from Greek National Hail Suppression Program (GNHSP) data. We use two groups of GNHSP data, namely, hailstorm data and seeding data. The following statements have being chosen as the goal of the study:

Statement 1: The prediction of the seeding parameters from the storm characteristics.
Statement 2: The determination of the storm type from the seeding parameters.
Statement 3: The determination of the required seeding time parameter from those parameters that describe the actual seeding.

The Greek Agricultural Insurance Organization applies the Greek National Hail Suppression Program by airborne seeding in order to minimize crop damages caused by hailstorms. The GNHSP project includes geographical regions of Northern and Central Greece. It is categorized as a weather modification program, based on the so-called “beneficial competition conceptual model”, and it is implemented by real time seeding of storms that are potential candidates for hail creation. The seeding material consists of silver iodide (AgJ), which is injected in the storm using suitably modified aircrafts [6]. The flights are monitored and

controlled by a meteorological radar control unit that detects and records hailstorms on a continuous basis.

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

2. Dataset used for Data Mining

The data recorded by the Thunderstorm Identification, Tracking, Analysis and Nowcasting system (TITAN) [3], 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 [5]. The data were recorded during the storm activity from April to September 2005 in the two protected areas of Northern (Macedonia) and Central (Thessaly) Greece. The SCC structured data represent the values of the SCC attributes, and, more specifically, the *Area of appearance*, *Type*, *New growth zone*, *Life time*, *Reflectivity*, *Cloud top*, *Speed*, and *Required seeding time*.

The structure of cloud systems and their classification in different categories follows the classification of SCC [5], [2]. 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. Regarding the values of *Area of appearance*, “M” is used for the project area of Macedonia and “L” for the project area of Thessaly. The possible values of the *New growth zone* attribute are the eight classes of azimuth orientations, e.g., the northeast site of the storm is represented by “NE”, and they indicate the appropriate locations for seeding penetrations. *Life time* denotes the storm life time in minutes, *Reflectivity* the storm intensity in dBZ, *Cloud top* the height of the cloud top above the ground level in Km, *Speed* the speed of the storm in Km/h, and, *Required seeding time* the required duration of storm seeding time in minutes according to GNHSP seeding criteria. Table 1 shows the statistics for the five continuous numerical attributes mentioned above, and Table 2 the categories we obtained for each attribute by transforming them to ordinal (categorical) data during the pre-processing phase. For the transformation we used equal frequencies and the chosen number of categories for each attribute was based on the special meaning of the GNHSP data and their potential contribution to the goals of the study.

	Life time (min)	*Reflectivity (dBZ)*	*Cloud top (Km)*	*Speed (Km/h)*	*Required seeding time (min)*
Mean	64.5	52	8.8	21.4	24.5
St. dev.	34.3	6.2	1.6	13	20.8
Min.	19	35	5.5	0	0
Max.	240	69	13	61	161

Table 1. Statistics of the storm parameters

Life time (min)	*Reflectivity*	*Cloud top (Km)*	*Speed*	*Required seeding*

	(dBZ)		*(Km/h)*	*time (min)*
19-38 (short)	35-45	5.5-7.5 (low)	0-12	0-12 (short)
39-48 (short-medium)	46-49	8-8.5 (low-medium)	13-20	13-26 (medium)
49-63 (medium)	50-51	9-9.5 (medium-high)	21-30	27-161 (long)
64-81 (medium-long)	52-54	10-13 (high)	31-61	
82-240 (long)	55-58			
	59-69			

Table 2. Categories of the storm parameters

The seeding data consists of the *Seeding material mass consumption* that is the amount of material in grams used per storm, the *Seeding time* that is the actual seeding duration time in minutes per storm, and, the *Seeding rate* that expresses the mean seeding rate per storm in grams per second. Again, during the pre-processing phase these continuous numerical attributes were transformed to ordinal (categorical) ones. The statistics of the original numerical attributes are shown in Table 3. The three categories that best represent the data and were obtained using equal frequencies are shown in Table 4.

	Seeding material mass consumption (g)	*Seeding time (min)*	*Seeding rate (g/s)*
Mean	1606.5	22.1	1.23
St. dev.	1538	19.2	0.59
Min.	20	3	0.03
Max.	8520	94	3.43

Table 3. Statistics of the seeding parameters

Seeding material mass consumption (g)		*Seeding time (min)*		*Seeding rate (g/s)*	
20-670	small	3-10	short	0.03-0.89	low
671-1810	medium	11-25	medium	0.90-1.41	medium
1811-8520	large	26-94	long	1.42-3.43	high

Table 4. Categories of the seeding parameters

The seeding data were calculated for each storm by a database application using imported data provided by the RDTS (Radar Data Telemetry System) data acquisition system of the aircrafts [6].

3. Methodology

The aim of the study is (a) the prediction of the seeding parameters from the storm characteristics, (b) the determination of the storm type from the seeding parameters, and, (c) the determination of the required seeding time parameter from

the actual seeding parameters. The storm parameters that we presented in the previous section were chosen in order to examine their contribution to the prediction of the seeding parameters. The potential of having the real meteorological radar data predict the seeding parameters for a storm is invaluable during seeding operations and can contribute to the efficient management of resources. Furthermore, it may be very useful to be able to determine and verify the type of each seeded storm from the seeding parameters, or in other words, to use the generated rules as an alternative way for performing meteorological radar image analysis. Finally, by using the actual seeding parameters to determine the required seeding time and examining the difference between the actual and required seeding times we can have a great means for evaluating the seeding operations.

The examined statements and the two groups of datasets we have chosen contain attributes that play multiple roles in the data mining process. Hence, the association rule mining techniques were chosen to discover interesting associations between those attributes. Association rules are unlike traditional classification rules in that an attribute appearing as a precondition in one rule may appear in the consequent of a rule. Association rule generators allow the consequent of a rule to contain one or several attribute values.

Association rules can be generated using a traditional approach, however, when several attributes are present, this process becomes unreasonable because of the large number of possible conditions for the consequent of each rule. Special algorithms have been developed to generate association rules efficiently. One such algorithm is the Apriori algorithm [1]. This algorithm generates item sets, which are attribute-value combinations that meet a specified coverage requirement. Those attribute-value combinations that do not meet the coverage requirement are discarded. Because of this, the rule generation process can be completed in a reasonable amount of time. Apriori association rule generation is a two-step process. The first step is item set generation. The second step uses the generated item sets to create a set of association rules [4].

4. Analysis and results

During the GNHSP operations, the aircrafts fly and perform seeding runs in the hailstorms that appear inside the project areas. The respective operational data from April to September 2005 were stored in a relational database. Next, we created a table of 241 records, one for each storm and its associated seeding details. Next, we imported this data in the WEKA data mining tool [7]. In the WEKA environment, we performed a preprocessing step to the target data (see Section 2), and, then we proceeded to the model construction by applying the Apriori algorithm to the given dataset. The chosen minimum values for Confidence and Support were set to 0.7 and 0.1 respectively. In addition, a second run was performed decreasing the Support lower bound to almost 0 and increasing the minimum Confidence to 0.8, in order to find out the rare but very interesting storm type values of “SU” and “L” in the antecedent and consequent of the generated rules.

The result was a large set of rules that associated the storm and seeding parameters. In order to give answers to the three statements that are the objective of this paper and to make easier and more effective the interpretation step, the output was grouped first by statement (the rules which refer to a given statement), then by attribute or combinations of attributes that appear in the consequent of a

rule, and, finally, by attribute or combinations of attributes that appear in the antecedent of a rule.

The first run gave a set of 389 rules from which we chose 84 rules that refer to the three statements, with a distribution of 44 rules for statement 1, 25 for statement 2 and 15 for statement 3. Also, the second run added 7 more rules to statement 2 and 8 to statement 3, increasing the output to 99 rules.

4.1 Rules for Statement 1

Regarding statement 1, the first run generated 44 rules with 15 of them having in the consequent the seeding parameter *Seeding material mass consumption*, 23 the *Seeding time* and 6 both the *Seeding material mass consumption* and the *Seeding time*. It is mentioned that the second run with almost zero Support did not produce any extra rules for statement 1. Also, the seeding parameter *Seeding rate* did not appear in the consequent of the rules. The most interesting rules for statement 1 are shown in Table 5, where in the antecedent of a rule may appear the storm parameters *Life time, Required seeding time, Cloud top, Type* and *New growth zone*, while in the consequent the seeding parameters *Seeding material mass consumption, Seeding time* and *Seeding rate*.

Life time	*Required seeding time*	*Cloud top*	*Type*	*New growth zone*	➔	*Seeding material mass consumption*	*Seeding time*	*CONF.*
short					➔	small		0.7
short	short				➔	small		0.76
	short		S		➔	small		0.73
	long				➔	large		0.73
	long		M		➔	large		0.76
long	long				➔	large		0.76
long		high			➔	large		0.75
	long	high			➔	large		0.82
	long	high	M		➔	large		0.84
short					➔		short	0.75
short	short				➔		short	0.81
	short				➔		short	0.71
	short		S		➔		short	0.75
	short	low			➔		short	0.79
	short			NE	➔		short	0.83
		low	S		➔		short	0.7
long					➔		long	0.71
long			M		➔		long	0.76
long		high			➔		long	0.81
	long				➔		long	0.76
	long		M		➔		long	0.79
	long	high			➔		long	0.82
long	long				➔		long	0.83
	long	high	M		➔		long	0.84
long	long	high			➔		long	0.86

long	long		M		➔		long	0.86
	long	high			➔	large	long	0.76
long	long				➔	large	long	0.73
	long		M		➔	large	long	0.7

Table 5: Prediction of seeding parameters from storm parameters.

Describing some of the important rules for statement 1, we remark that a storm with a short *Life time* (up to 38 min) has a treatment of a small *Seeding material mass consumption* (up to 670 g) with an accuracy level of 0.7. Also, a storm with a long *Required seeding time* (greater than 26 min) has a treatment of a large *Seeding material mass consumption* (greater than 1810g) with an accuracy level of 0.73. A short *Life time* (up to 38 min) leads to a short *Seeding time* (up to 10 min) with an accuracy level of 0.75, and a storm with a low *Cloud top* (up to 7.5 Km) and *Type* "S" has a short *Seeding time* with an accuracy level of 0.7. A storm with a long *Life time* (greater than 81 min), a long *Required seeding time* and a high *Cloud top* (greater than 9.5 Km) has a long *Seeding time* (greater than 25 min) with an accuracy level of 0.86.

4.2 Rules for Statement 2

Regarding statement 2, the first run generated 21 rules and the second run 7 more rules having in the consequent the storm *Type* "M". Also, the first run generated 4 rules having in the consequent the storm *Type* "S", while the very low frequency storm *Type* "SU" and "L" did not appear at all. The most interesting rules for statement 2 are shown in Table 6, where in the antecedent of a rule may appear the seeding parameters *Seeding time, Seeding material mass consumption* and *Seeding rate* and in the consequent the storm parameter *Type*.

Seeding time	*Seeding material mass consumption*	*Seeding rate*	➔	*Type*	*CONF.*
	large		➔	M	0.91
long			➔	M	0.85
		high	➔	M	0.7
	large	high	➔	M	0.92
long		medium	➔	M	0.88
long		high	➔	M	0.88
medium	large		➔	M	1
short		low	➔	S	0.71
short	small	low	➔	S	0.71
short	small		➔	S	0.7

Table 6: Determination of the storm type from the actual seeding parameters.

Describing some of the important rules for statement 2, we remark that a large *Seeding material mass consumption* leads to the most frequent storm *Type* "M" with an accuracy level of 0.91, and combined with a medium *Seeding time* (between 11 and 25 min) leads again to *Type* "M" with the highest accuracy level of 1. A short

Seeding time and a low *Seeding rate* (less than 0.9 g/s) leads to *Type* “S” with an accuracy level of 0.71.

4.3 Rules for Statement 3

Regarding statement 3, the first run generated 15 rules having in the consequent a short (6 rules) and a long (9 rules) *Required seeding time*, and the second run generated 8 more rules having in the consequent a medium *Required seeding time*. The most interesting rules for statement 3 are shown in Table 7, where in the antecedent of a rule may appear the seeding parameters *Seeding time, Seeding material mass consumption* and *Seeding rate* and in the consequent the *Required seeding time*.

Seeding time	*Seeding material mass consumption*	*Seeding rate*	➔	*Required seeding time*	*CONF.*
	large		➔	long	0.72
	large	medium	➔	long	0.79
	large	low	➔	long	0.89
long			➔	long	0.78
long		medium	➔	long	0.81
long		high	➔	long	0.83
long	large		➔	long	0.84
long	large	low	➔	long	0.89
long	small		➔	medium	1
long	medium	medium	➔	medium	1
short	small		➔	short	0.72
short		low	➔	short	0.71

Table 7: Determination of the Required seeding time from the actual seeding parameters.

Describing some of the above important rules, a large *Seeding material mass consumption* and a low *Seeding rate* lead to a long *Required seeding time* having an accuracy level of 0.89. Also, a long *Seeding time* and a small *Seeding material mass consumption* lead to a medium *Required seeding time* (between 13 and 26 min) having the highest accuracy level of 1, whereas, a short *Seeding time* and low *Seeding rate* lead to a short *Required seeding time* having an accuracy level of 0.71.

5. Conclusion

The last stage of a KDD process and the most crucial is the output interpretation. In our case, there are three groups of chosen rules that potentially answer the three statements we defined as the goal of the study. Concerning the rules for statement 1, we could remark that the main storm characteristics that can determine the seeding parameters are the *Life time* and the *Required seeding time*. *Required seeding time* physically is a portion of the *Life time* since the role of *Life time* is very important. In addition, *Cloud top* and *Type* contribute in many cases and *New growth zone* rarely in increasing the accuracy of a rule. Also, a spatial characteristic, the *Area of appearance* appears many times in the antecedent of a rule but its

contribution is only secondary in improving the accuracy of a rule (i.e., there is no differentiation in the characteristics of storms in Macedonia and Thessaly). On the contrary, the storm characteristics *Reflectivity* and *Speed* never appear in the antecedent of the important rules. Only two of the three seeding parameters could be determined, *Seeding material mass consumption* and *Seeding time*, and specifically the short and long values of the latter and the small and large values of the former. *Seeding rate* does not appear in the consequent of an important rule and the medium values for both *Seeding material mass consumption* and *Seeding time* is not predicted. The accuracy varies between 0.7 and 0.86.

The consequent of statement 2 is the storm *Type* and the rules determine the most frequent type “M” and the relatively frequent type “S”, but not the rare types (frequencies less than 1%) “SU” and “L”. The accuracy for *Type* “M” reaches in some cases the highest level of 1 and for *Type* “S” 0.71. All the seeding parameters could determine *Type* “M” and “S” and it seems that the most accurate is the *Seeding material mass consumption* followed by the *Seeding time*. The consequent of statement 3 is the *Required seeding time* and the rules can determine all its potential values. In the antecedent all the seeding parameters contribute to the determination with the *Seeding rate* having a secondary role. The accuracy reaches in some cases the highest level of 1.

Taking into account that there are accurate and simple rules with one or two attributes in the antecedent, it is believed that the rules presented in Tables 5, 6 and 7 are easy to use and have a practical value. They could contribute to the management of resources, the meteorological radar image analysis, and the seeding operations evaluation.

References

[1] Agrawal R., Imielinski T., Swami A., “Mining Association Rules between sets of items in large Databases”, In P. Buneman and S. Jajordia, eds., *Proceedings of the ACM Sigmoid International Conference on Management of Data,* New York: ACM, 1993.

[2] Browning K. A., “The structure and mechanisms of hailstorms”, Meteor*. Monogr*, vol. 38, pp. 1-43, 1977.

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

[4] Roiger R. J., Geatz M. W., “Data Mining: A tutorial based primer”, Pearson Education, Inc., 2003.

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

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

[7] Witten I. H., Frank E., “Data Mining: Practical Machine Learning Tools and Techniques”, 2nd Edition, Morgan Kaufmann Publishers, June 2005.