Extraction of the Convective Day Category Index using Data Mining Techniques

Evangelos G. Tsagalidis [1)], Leonidas Karamitopoulos [2)],
Georgios Evangelidis [2)], Dimitris A. Dervos [3)],
1) Hellenic Agricultural Insurance Organization, Meteorological Applications Center,
International Airport Macedonia, GR-55103 Thessaloniki, GREECE, vang@the.forthnet.gr
2) Department of Applied Informatics, University of Macedonia,
156 Egnatia St., GR-54006, Thessaloniki, GREECE, {gevan, lkaramit}@uom.gr
3) Information Technology Dept., T.E.I.,
P.O. BOX 141, GR-57400 Sindos, GREECE, dad@it.teithe.gr

Abstract: *One of the tasks of the Hellenic Hail Suppression Program is the determination of the observed Convective Day Category (CDC) Index. This process is accomplished by having the meteorologists analyze the operational data manually. To automate and speed up this procedure we have developed an application in the CLIPS Expert System environment that calculates the observed CDC index. In this paper we examine the appropriate data mining techniques that could be used to extract this index from operational data automatically.*

Keywords: hail suppression program, convective day category index, data mining, classification, decision trees

1. INTRODUCTION

The Meteorological Applications Center (KEME) is the section of Hellenic Agricultural Insurance Organization (ELGA) that applies the Hellenic Hail Suppression Program (HHSP) in two project areas of northern and central Greece. The aim of the HHSP is the protection of the cultivation from damages due to hail. The program is a Weather Modification Program and is applied via cloud seeding based on the conceptual model of "beneficial competition" of the field of cloud microphysics. From April up to September, specially equipped aircrafts guided by meteorological radars, approach at the appropriate time the appropriate places of a candidate cloud for hail creation and perform seeding. The seeding material is artificial ice nuclei (AgJ) that competes the physical nuclei for collecting the "liquid water content" of the cloud and, as a result, reduces the hail size in the cloud and increases the possibility that the hailstone will melt during its fall under the cloud base.

Every day (from 09:00 UTC up to 09:00 UTC the next day) a meteorological forecast is performed for the convective activity around the project areas, which categorizes the day with the use of the Forecasted CDC (Convective Day Category) index [1]. The CDC index follows a range of 9 categories, using a number from -3 up to +5. Describing briefly the meaning of the values of the CDC index, -3 corresponds to no convection, -2 to shallow convective cloudiness, -1 to rain shower, 0 to thunderstorm but without hail on the ground, +1 small hail of pea size on the ground, +2 hail of grape size, +3 hail of walnut size, +4 hail of golf ball size, and +5 hail larger than the golf ball size. The Forecasted CDC in the Hail Suppression Program increases the level of operational readiness and improves the management of the resources.

The meteorologists of the Hail Suppression Program determine the observed CDC for a particular day from data extracted from the available recorded data for that day. The observed CDC is determined from the highest recorded CDC value during that day, according to criteria fulfilled in the following order:

a) measurements of hail from the Hailpad Network,
b) damage reports on the cultivation,
c) radar echo reflectivity from clouds,
d) special echo patterns on the Plan Position Indicator (PPI) and Range Height Indicator (RHI) of the meteorological radar,
e) height of the cloud top measured by the radar, and
f) visual and sound observations.

The observed CDC is compared to the Forecasted CDC in order to evaluate the forecast. Also, the observed CDC is considered to be a significant parameter that combined with other data contributes to the studies and the research associated with the improvement of the application and the operation of the Hail Suppression Program.

In the present study, we present an application we have developed in the CLIPS Expert System environment [2] to calculate the observed CDC index and we use the recorded data of one period of the HHSP Program to examine the effectiveness of various data mining techniques [3] in deriving the CDC index. Section 2 presents the CLIPS application. Section 3 describes the dataset we used for applying the data mining algorithms that we describe in Section 4. In Section 5 we present the results we obtained by experimenting with the chosen algorithm and finally we conclude the paper in Section 6.

2. EXPERT SYSTEM APPLICATION

CLIPS is a computer programming language designed for writing applications called expert systems. An expert system is a special program intended to model human expertise or knowledge. The model we developed in the CLIPS Expert System environment is based on the facts and the rules that correspond to the CDC index of HHSP. The system interacts with the user through a series of questions and depending on the provided answers a

certain final result appears on the screen accompanied by an appropriate message. The output messages are related to the input data and guide the user in the case he/she provides inconsistent data. The system is a data-driven program and, hence, it is considered to be a rule-based expert system. The priorities of the chosen criteria that determine the observed CDC are assured with the use of appropriate flow control commands. This approach results in the development of an objective system.

Table 1 shows the names, values, and definitions of the variables that were used in the program.

Table 1. Variables in the Expert System Application

Variable	Values	Definition
hailnet	yes, no	Hail recorded in the Hail Network
hailsize	1, 2, 3, 4, 5	The recorded maximum hailstone in the network
hailreport	yes, no	Valid report for hail out of the network
reflectivity	1, 2, 3, 4, 5, 0, -1, -2, -3	Radar echo reflectivity of the storms
h-trop+15	yes, no	The height of the cloud top is 1,5 Km above the tropopause
h-trop+1	yes, no	The height of the cloud top is 1 Km above the tropopause
h-trop+05	yes, no	The height of the cloud top is 0,5 Km above the tropopause
h-trop	yes, no	The height of the cloud top reaches the tropopause
ppirhi-severe	yes, no	Severe storm patterns on radar
ppirhi-intense	yes, no	Intense storm patterns on radar
rhi-wer	yes, no	Weak echo vault or WER on RHI screen
h-30	yes, no	The height of the cloud top is above the level of -30^0C
h30	yes, no	The height of the cloud top reaches the level of -30^0C
h-12	yes, no	The height of the cloud top is above the level of -12^0C
dh-3	yes, no	The cloud depth is higher than 3 Km
dh3	yes, no	The cloud depth is lower than 3 Km

The first question asked by the system is about the recorded hail in the network. If there is hail in the network questions regarding its size and existence of valid reports for hail outside the network follow. Otherwise, the system asks for the value of reflectivity and depending on the answer provided by the user, the system also asks for the values of the cloud top and the storms patterns on the PPI and RHI screens of the radar.

Using the recorded data from the 169 operational days of year 2000 (usually the annual period of the Program is 169 days - from 15 April up to 30 September) we calculated the observed CDC index for each day.

3. DATASET USED FOR DATA MINING

Our goal is to use various data mining techniques in order to classify days according to their CDC index value. We use the observed CDC index we calculated in the previous section with CLIPS as a class variable, that is, given as input values for certain variables we want to build models that predict the observed CDC index for a particular day.

During the preprocessing phase, we made the appropriate transformations to the operational data. More specifically, we created four (4) variables, namely, *hailsr* that express the recorded hail size whether inside or outside the Hail Network (see Table 2), *reflectivity* that expresses the radar echo reflectivity (see Table 3), *ctop* that expresses the level of the cloud top (see Table 4), *ppirhi* that expresses the storm patterns on the radar screens (see Table 5), and finally, *calculated CDC* which is the CDC provided by the expert system.

Table 2. Values of the input variable *hailsr*

Value	Description
0	No hail.
1	Hail in the network of category size 1.
2	Hail in the network of category size 2.
3	Hail in the network of category size 3.
4	Hail in the network of category size 4.
5	Hail in the network of category size 5.
6	No hail in the network, hail report outside.

Table 3. Values of the input variable *reflectivity*

Value	Description
-3	Radar echo reflectivity of category -3.
-2	Radar echo reflectivity of category -2.
-1	Radar echo reflectivity of category -1.
0	Radar echo reflectivity of category 0.
1	Radar echo reflectivity of category 1.
2	Radar echo reflectivity of category 2.
3	Radar echo reflectivity of category 3.
4	Radar echo reflectivity of category 4.
5	Radar echo reflectivity of category 5.

Table 4. Values of the input variable *ctop*

Value	Description
-3	Cloud top of category -3.
-2	Cloud top of category -2.
-1	Cloud top of category -1.
0	Cloud top of category 0.
1	Cloud top of category 1.
2	Cloud top of category 2.
3	Cloud top of category 3.
4	Cloud top of category 4.
5	Cloud top of category 5.

Table 5. Values of the input variable *ppirhi*

Value	Description
3	Pattern of category 3.
4	Pattern of category 4.
5	Pattern of category 5.

Our dataset did not contain any +3, +4 and +5 values for the *calculated CDC*, so the input variable *ppirhi* did not have any recorded values. Thus, we decided to exclude this variable from our analysis.

4. METHODOLOGY

The problem of extracting the observed CDC index from our operational database is a typical classification problem. Formally, the classification problem is stated as below:

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 three variables (hail size, reflectivity and cloud top) and the values of the observed CDC index correspond to the classes. There is a wealth of classification techniques mainly developed from the fields of Statistics and Machine Learning, such as regression, Bayesian classification, nearest neighbor, decision trees, neural networks, and support vector machines.

We chose decision tree-based algorithms to perform classification in our dataset, mainly because of their two basic characteristics of these algorithms, namely, interpretability and facility in generating rules that are desirable from the end users [4]. Furthermore, two of the main disadvantages of the decision-tree technique, i.e., overfitting and the possibility of producing a large tree that needs pruning, do not apply in our case since the size of our dataset is small and accurately represents a typical season with respect to the CDC index. In other words, this means that the derived decision tree is expected to be adequately small and the problem of overfitting will not appear since the variability of the observed CDC index is fairly small. For the same reason, speed and scalability (time to construct the model, time to use the model and efficiency in disk-resident databases) are not important issues. The three most popular decision tree algorithms we chose to test on our dataset are: CART, C4.5 and CHAID.

The CART algorithm of L. Breiman et al. [5] has been a staple of machine learning experiments [6]. CART builds a binary tree by splitting the records at each node according to a function of a single input variable. The measure used to evaluate the best splitter is Gini. Splits are found in order to maximize the homogeneity of child nodes with respect to the value of the dependent variable. Gini is based on squared probabilities of membership for each category of the dependent variable. It reaches its minimum (zero) when all tuples in a node fall into a single category.

In our experiments, the minimum change in improvement, that is, the minimum decrease in impurity required to split a node has been set to 0.0001, and the minimum number of tuples in a parent or child node has been set to 2. Although this number is very small, it does not affect the generality of the tree, since our dataset can be considered to be typical for any year. The maximum number of levels has been set to 5 and the maximum number of surrogates has been set to 2. Surrogates are used to classify tuples that have missing values on independent variables used in the tree [7].

C4.5 is a decision-tree algorithm that J. R. Quinlan has been evolving and refining for many years [8, 9]. C4.5 is very similar to CART. One difference is that C4.5 produces trees that are not necessarily binary. The splitting criterion is based on the concept of information gain and the corresponding measure is the entropy or information.

In our experiments, the minimum number of tuples in a parent or child node has been set to 2.

CHAID algorithm [10] grows a tree by splitting the tuples at each node according to the statistically significant differences that are produced in the target variable values. The measure used to evaluate the best splitter is Pearson's Chi-square. CHAID is restricted to categorical or ordinal variables and attempts to stop growing the tree before overfitting occurs.

In our experiments, the significance level for splitting nodes and merging categories has been set to 0.05 and adjusted using the Bonferroni method. The minimum number of tuples in a parent or child node has been set to 2. The maximum number of levels has been set to 3.

In all algorithms, missing values of nominal independent variables have been treated as missing values. Finally, all three algorithms are applied in two modes, (a) using the whole sample as the training set, and (b) validating the method by using half the sample as the training set and the rest half of it as the testing sample.

SPSS [7] was used to apply CART (CRT) and CHAID and WEKA [11] to apply C4.5 (J48) on our dataset. In the following we will use the notation CARTa and CARTb (accordingly CHAIDa, CHAIDb, C4.5a, C4.5b) to refer to the two modes of application of the algorithms.

5. ANALYSIS AND RESULTS

In this section we analyze the outcomes of the application of the three data mining algorithms on our dataset.

First, we have to make the remark that only C4.5 gave exactly the same decision tree of depth 2 (see Fig.2 in the Appendix) for both modes of application (C4.5a and C4.5b). The decision trees of CARTa (see Fig.1 in the Appendix) and CARTb had the highest depth, 5 and 4 respectively, whereas the decision trees of CHAIDa (see Fig.3 in the Appendix) and CHAIDb had a depth of 3 and 2, respectively.

As expected, the risk estimate of C4.5a is lower than the one of C4.5b (0,6% vs. 8,2%). Both CARTa and CHAIDa had a risk estimate of 2,4%, whereas the CARTb and CHAIDb figures for the risk estimate were 7,4% and 4,5% respectively.

From the expert's (meteorologist's) point of view the derived decision trees are quite satisfactory classification models for the observed CDC index. The low depth of the decision trees was expected due to the small number of the input variables; however, these variables are adequate for predicting CDC. Although the size of the data set is small, it covers the hail period in the protected areas of HHSP. Furthermore, this data set could be considered very close to a typical annual distribution and can be considered as a representative sample. We also mention that CDC values of +3, +4 and +5 refer to extreme cases for the reference area (especially the +4 and +5 values), so the missing records for these values do not affect the

results.

Regarding CARTa, we obtain the best results for classes -2, -1, and 0 (see Table 6). For class –3, 1 instance (1,1%) is misclassified to class -2, for class +1, 1 instance (7,7%) is misclassified to class 0, and, for class +2, 2 instances (40%) are misclassified to class +1.

Table 6. Confusion Matrix - CARTa

Classification

Observed	Predicted -3.00	-2.00	-1.00	.00	1.00	2.00	Percent Correct
-3.00	92	1	0	0	0	0	98.9%
-2.00	0	4	0	0	0	0	100.0%
-1.00	0	0	11	0	0	0	100.0%
.00	0	0	0	43	0	0	100.0%
1.00	0	0	0	1	12	0	92.3%
2.00	0	0	0	0	2	3	60.0%
Overall Percentage	54.4%	3.0%	6.5%	26.0%	8.3%	1.8%	97.6%

Growing Method: CRT
Dependent Variable: clips

Regarding CARTb, we obtain the best results for classes -2, and -1 (see Table 7). For class –3, 1 instance (1,8%) is misclassified to class -2, for class 0, 4 instances (19%) are misclassified to classes +1 and +2, for class +1, 1 instance (14,3%) is misclassified to class +2, and, for class +2, 1 instance (33,3%) is misclassified to class +1.

Table 7. Confusion Matrix - CARTb

Classification

Sample	Observed	Predicted -3.00	-2.00	-1.00	.00	1.00	2.00	Percent Correct
Training	-3.00	38	0	0	0	0	0	100.0%
	-2.00	0	2	0	0	0	0	100.0%
	-1.00	0	0	5	0	0	0	100.0%
	.00	0	0	0	20	2	0	90.9%
	1.00	0	0	0	0	5	1	83.3%
	2.00	0	0	0	0	1	1	50.0%
	Overall Percentage	50.7%	2.7%	6.7%	26.7%	10.7%	2.7%	94.7%
Test	-3.00	54	1	0	0	0	0	98.2%
	-2.00	0	2	0	0	0	0	100.0%
	-1.00	0	0	6	0	0	0	100.0%
	.00	0	0	0	17	2	2	81.0%
	1.00	0	0	0	0	6	1	85.7%
	2.00	0	0	0	0	1	2	66.7%
	Overall Percentage	57.4%	3.2%	6.4%	18.1%	9.6%	5.3%	92.6%

Growing Method: CRT
Dependent Variable: clips

CHAIDa behaves exactly like CARTa (we do not include the corresponding confusion matrix since it is the same as the one in Table 6).

Regarding CHAIDb, we obtain the best results for classes -3, -2, -1, and 0 (see Table 8). For class +1, 1 instance (12,5%) is misclassified to class 0, and, for class +2, 3 instances (100%) are misclassified to class +1.

Table 8. Confusion Matrix - CHAIDb

Classification

Sample	Observed	Predicted -3.00	-2.00	-1.00	.00	1.00	2.00	Percent Correct
Training	-3.00	45	1	0	0	0	0	97.8%
	-2.00	0	3	0	0	0	0	100.0%
	-1.00	0	0	4	0	0	0	100.0%
	.00	0	0	0	21	0	0	100.0%
	1.00	0	0	0	0	5	0	100.0%
	2.00	0	0	0	0	2	0	.0%
	Overall Percentage	55.6%	4.9%	4.9%	25.9%	8.6%	.0%	96.3%
Test	-3.00	47	0	0	0	0	0	100.0%
	-2.00	0	1	0	0	0	0	100.0%
	-1.00	0	0	7	0	0	0	100.0%
	.00	0	0	0	22	0	0	100.0%
	1.00	0	0	0	1	7	0	87.5%
	2.00	0	0	0	0	3	0	.0%
	Overall Percentage	53.4%	1.1%	8.0%	26.1%	11.4%	.0%	95.5%

Growing Method: CHAID
Dependent Variable: clips

C4.5a gave best results for all classes, except for class –3 where 1 instance (1,1%) is misclassified to class –2 (see Table 9). C4.5b gave best results for class -1, while for class –3, 1 instance (2,1%) is misclassified to class -2, for class 0, 2 instances (8,3%) are misclassified to class +2, for class +1, 2 instances (33,3%) are misclassified to class +2, and finally, for class +2, 2 instances (66,6%) are misclassified to class +1 (see Table 10).

Table 9. Confusion Matrix - C4.5a

a	b	c	d	e	f	classified as
92	0	0	1	0	0	a = -3
0	43	0	0	0	0	b = 0
0	0	13	0	0	0	c = 1
0	0	0	4	0	0	d = -2
0	0	0	0	5	0	e = 2
0	0	0	0	0	11	f = -1

Table 10. Confusion Matrix - C4.5b

a	b	c	d	e	f	classified as
46	0	0	1	0	0	a = -3
0	22	0	0	2	0	b = 0
0	0	4	0	2	0	c = 1
0	0	0	1	0	0	d = -2
0	0	2	0	1	0	e = 2
0	0	0	0	0	4	f = -1

The rules that SPSS has created for CART and CHAID are quite interesting. Below, we include the rule derived from Node 6 of the CHAIDa decision tree:

```
Node 6
IF (reflecti = 1 OR reflecti = 2)  AND
   (hailsr != 1  AND  hailsr != 6  AND  hailsr != 2)
THEN
   Node = 6
   Prediction = 0
   Probability = 1.000000
```

Also, the rule derived from Node 7 of the CARTa decision tree is:

```
Node 7
IF (((reflecti = 0 OR reflecti = 1 OR reflecti = -2 OR
reflecti = 2 OR reflecti = -1) OR (reflecti != -3)  AND
((ctop = 2 OR ctop = 1 OR ctop = -1 OR ctop = -2 OR
ctop = 0) OR (ctop != -3)  AND  (hailsr = 1 OR hailsr =
6 OR hailsr = 2))))  AND  (reflecti != -1)  AND
(((hailsr = 0) OR (hailsr != 1  AND  hailsr != 6  AND
hailsr != 2)  AND  ((ctop = -3 OR ctop = -1 OR ctop = -
2 OR ctop = 0) OR (ctop != 2  AND  ctop != 1)  AND
(reflecti != 2))))  AND  (((reflecti = -2) OR (reflecti
!= 0  AND  reflecti != 1  AND  reflecti != 2)  AND
(ctop = -3 OR ctop = -2)))
THEN
   Node = 7
   Prediction = -2
   Probability = 0.800000
```

Based on our experience in deriving the observed CDC index for years, we claim that all the methods seem to give good results.

The algorithm of C4.5 builds a decision tree that uses only two independent variables (reflectivity and hailsr). The only problem is when reflectivity is 0 and hailsr is 2; then the resulting CDC index is 0, but it should be +2.

CARTa also uses the above-mentioned two variables (reflectivity and hailsr), while CARTb uses all three independent variables. When reflectivity is 0 and hailsr takes any value the resulting CDC index is 0, while it should be +1 or +2. CARTb appears to have a similar problem.

CHAIDa and CHAIDb use the reflectivity and the

hailsr variables. These methods appear to be the best ones, since they do not show any serious problem. However, the CHAIDb decision tree has a depth of 2 (one less than the corresponding CHAIDa decision tree).

6. CONCLUSION

In the present study we examined the possibility of applying data mining techniques on the operational data of the Hellenic Hail Suppression Program (HHSP) in order to extract the observed CDC index automatically. The resulting models are thought to be satisfactory with acceptable values of accuracy and could be used by meteorologists to speed up and automate the determination of the observed CDC index.

7. REFERENCES

1. E. G. Tsagalidis, D. B. Foris. EL.G.A. The Convective Day Category Index, *Annual Report of Hellenic Hail Suppression Program,* 1999.
2. J. C. Giarratano. *CLIPS User's Guide, version 6.10*. 1998.
3. M.H. Dunham. *Data Mining: Introductory and Advanced Topics*, Pearson Education, Inc. New Jersey, 2003.
4. S. K. Murthy. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery* 2(4) (1998). p. 345-389.
5. L. Breiman, J. Friedman, R. Olshen, C. Stone. *Classification and Regression Trees*. Wadsworth Int. Group, 1984.
6. M. J. A. Berry, G. Linoff. *Data Mining Techniques for Marketing, Sales, and Customer Support.* Wiley Computer Publishing, 1997.
7. SPSS. http://www.spss.com. Statistical Analysis Software.
8. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers. San Mateo, CA, 1993.
9. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1 (1986). P. 81-106.
10. G. V. Kass. An exploratory technique for investigating large quantities of categorical data, *Applied Statistics* 29(2) (1980). p. 119-127.
11. WEKA. http://www.cs.waikato.ac.nz/~ml/weka/. Data Mining with Open Source Machine Learning Software in Java.

APPENDIX

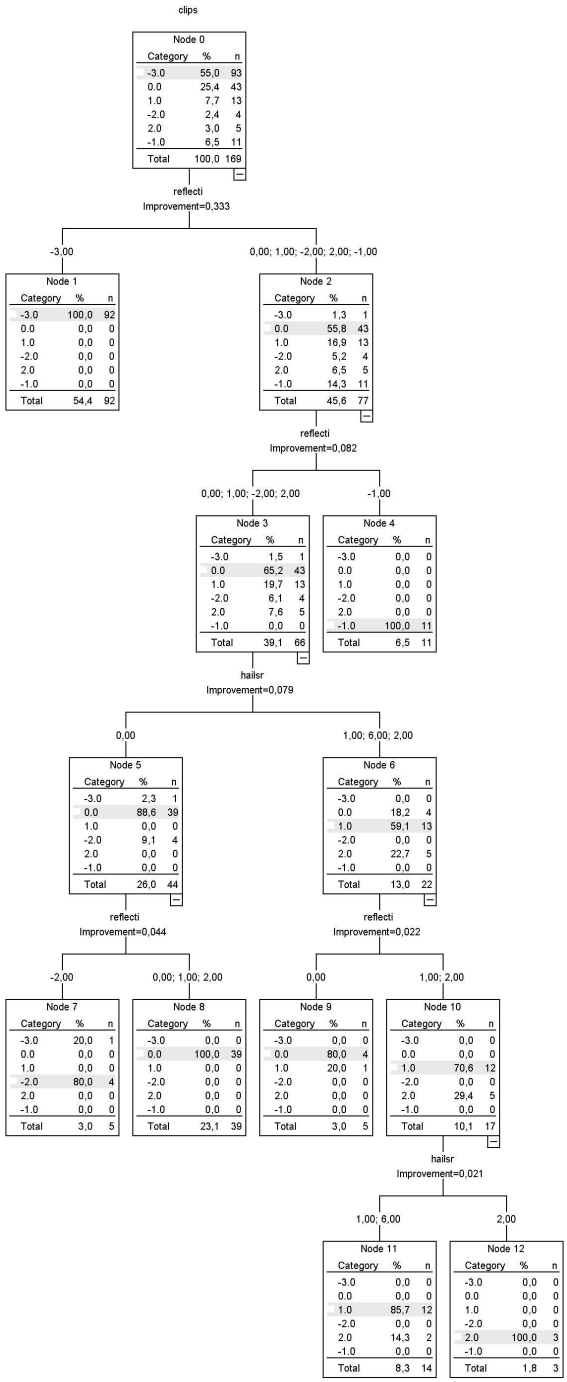

Fig.1 - CARTa decision tree

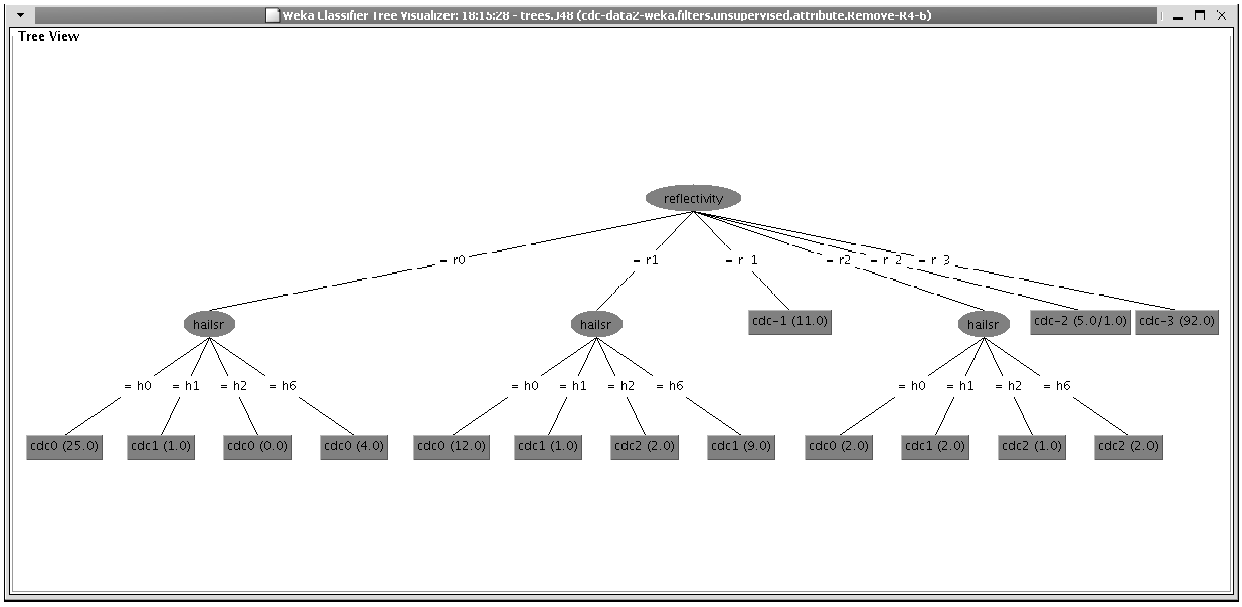

Fig.2 - C4.5 decision tree

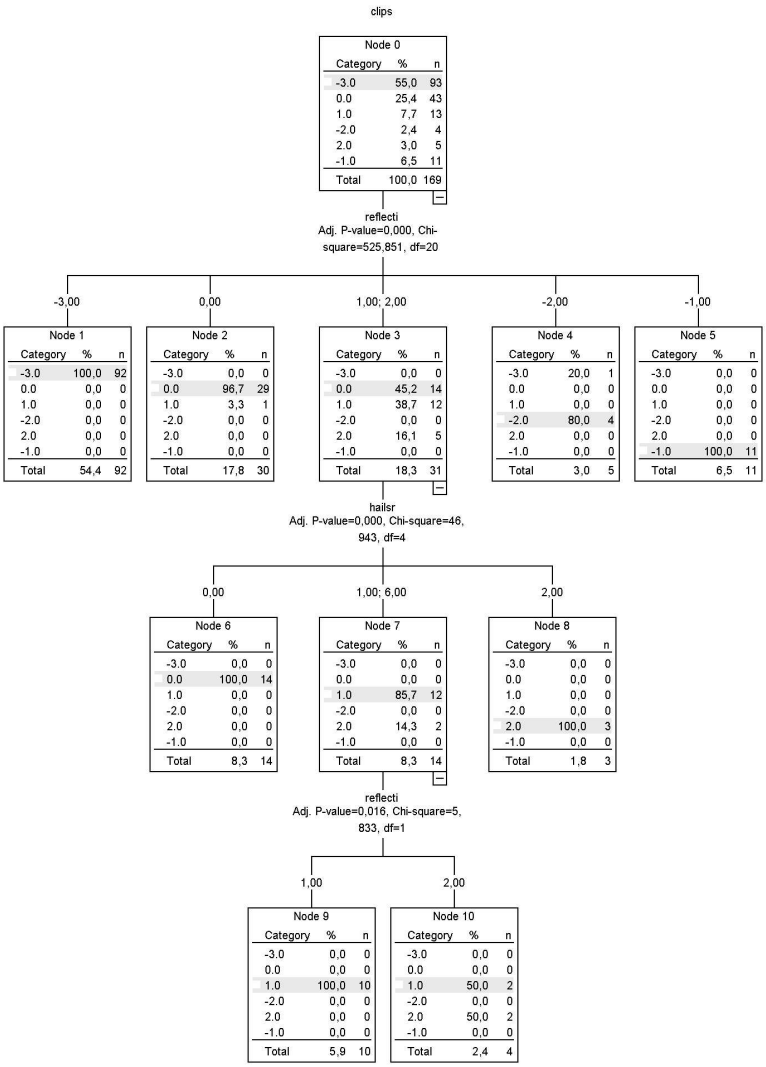

Fig.3 - CHAIDa desicion tree