BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXI No. 1/2009	77 - 82	Seria Matematică - Informatică - Fizică

A Decision Tree for Weather Prediction

Elia Georgiana Petre

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: elia_petre @yahoo.com

Abstract

A decision tree represents a decision support tool very often used because it is simple to understand and interpret. Classification and Regression Trees - CART - is a technique formed by a collection of rules based on values of certain variables in the modelling data set. This paper presents a small application of CART for whether prediction. It had been chosen the data collection registered over Hong Kong. The data was recorded between 2002 and 2005. To build the decision tree we used a free data mining software available under the GNU General Public License– Weka. Then, there are presented the decision tree, the results and the statistical information about the data used to generate the decision model.

Key words: *decision tree, CART algorithm, data mining, whether prediction*

Introduction

Since ancient times, weather prediction has been one of the most interesting and fascinating domain. The scientists have been trying to forecast the meteorological characteristics using a large set of methods, some of them more accurate than others. Lately, there has been discovered that data mining, a method developed recently, can be successfully applied in this domain.

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions [1]. The most commonly used techniques in data mining are: artificial neural networks, genetic algorithms, rule induction, nearest neighbour method and memory-based reasoning, logistic regression, discriminant analysis and decision trees.

In this paper we present how CART, one of the most popular decision tree algorithms, can be used in weather prediction domain.

Brief Description of Decision Trees

Decision trees models are commonly used in data mining to examine the data and to induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, and C5.0 [1].

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.

Depending on the algorithm, each node may have two or more branches. For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed this is called a multiway tree [2].

In this application, we use CART to build a classification tree to predict the temperature values around Hong Kong. For this, there are used meteorological data registered between 2002 and 2005. These aspects are to be discussed in more detail below.

Application Description

We would like to analyze meteorological data registered during the last years in the capital of China and thus try to forecast the future temperature values in Hong Kong.

In order to have a detailed outline of the weather parameters, we have used the data between 2002 and 2005 [4].

The data used to create our database include: *year, month, average pressure, relative humidity, clouds quantity, precipitations* and *average temperature*.

The next step of our application focuses on transforming these data in order to be used in Weka, a data mining specialized software. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning written in Java, developed at the University of Waikato. WEKA is a free software available under the GNU General Public License. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality [3].

Having in mind the fact that CART algorithm can only work with nominal variables, we have to adapt the data. With this intention we will make some transformations:

For *year* data, 2002 will be in Weka database 2, 2003 will be 3, 2004 is 4 and 2005 is 5.

The *month* parameter will be encoded like: 1 for January, 2 – February, 3 – March, 4 – April, 5 – May, 6 – June, 7 – July, 8 – August, 9 – September, 10 – October, 11 – November, 12 – December.

The *average pressure* values will be divided in equal intervals, each ones having attached a number: 1 – (1000, 1004), 2 – (1005, 1009), 3 – (1010, 1014), 4 – (1015, 1019), 5 – (1020, 1024), 6 – (1025, 1030).

The same codification is used for the *relative humidity* values: 1 – (60, 64), 2 – (65, 69), 3 – (70, 74), 4 – (75, 79), 5 – (80, 84), 6 – (85, 89), 7 – (90, 94), as well as for the *clouds quantity*: 1 – (30, 39), 2 – (40, 49), 3 – (50, 59), 4 – (60, 69), 5 – (70, 79), 6 – (80, 89), 7 – (90, 99) and for *precipitations*: 1 – (0, 99), 2 – (100, 199), 3 – (200, 299), 4 – (300, 399), 5 – (400, 499), 6 – (500, 599), 7 – (600, 699), 8 – (700, 799), 9 – (800, 899), 10 – (900, 999), 11 – (1000, 1999), 12– (2000, 2999), 13 – (3000, 3999).

The predicted variable, *average temperature* is measured in Celsius degrees and will be implemented in Weka using 1 for the (10, 14) interval, 2 for (15, 19), 3 for (20, 24), 4 for (25, 29) and 5 for (30, 34).

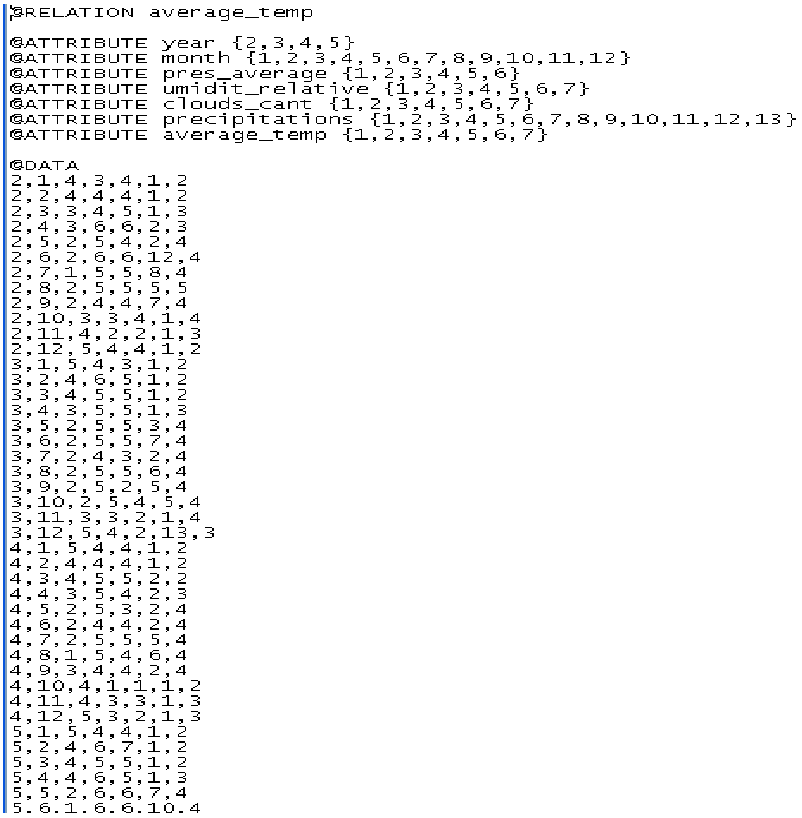
In figure 1 is presented file *weather.arff*, the Weka database after the transformation.

Running the Application

After this step, in order to generate the decision tree based on the CART algorithm in Weka, we must load our database first, followed by the proper selection of the algorithm from the list provided by this software and then press the Start button.

The output generated has the next sections:

- Run information
- Classifier model
- Evaluation on training set
- Detailed Accuracy By Class



```
@RELATION average_temp

@ATTRIBUTE year {2,3,4,5}
@ATTRIBUTE month {1,2,3,4,5,6,7,8,9,10,11,12}
@ATTRIBUTE pres_average {1,2,3,4,5,6}
@ATTRIBUTE umidit_relative {1,2,3,4,5,6,7}
@ATTRIBUTE clouds_cant {1,2,3,4,5,6,7}
@ATTRIBUTE precipitations {1,2,3,4,5,6,7,8,9,10,11,12,13}
@ATTRIBUTE average_temp {1,2,3,4,5,6,7}

@DATA
2,1,4,3,4,1,2
2,2,4,4,4,1,2
2,3,3,4,5,1,3
2,4,3,6,6,2,3
2,5,2,5,4,2,4
2,6,2,6,6,12,4
2,7,1,5,5,8,4
2,8,2,5,5,5,5
2,9,2,4,4,7,4
2,10,3,3,4,1,4
2,11,4,2,2,1,3
2,12,5,4,4,1,2
3,1,5,4,3,1,2
3,2,4,6,5,1,2
3,3,4,5,5,1,2
3,4,3,5,5,1,3
3,5,2,5,5,3,4
3,6,2,5,5,7,4
3,7,2,4,3,2,4
3,8,2,5,5,6,4
3,9,2,5,2,5,4
3,10,2,5,4,5,4
3,11,3,3,2,1,4
3,12,5,4,2,13,3
4,1,5,4,4,1,2
4,2,4,4,4,1,2
4,3,4,5,5,2,2
4,4,3,5,4,2,3
4,5,2,5,3,2,4
4,6,2,4,4,2,4
4,7,2,5,5,5,4
4,8,1,5,4,6,4
4,9,3,4,4,2,4
4,10,4,1,1,1,2
4,11,4,3,3,1,3
4,12,5,3,2,1,3
5,1,5,4,4,1,2
5,2,4,6,7,1,2
5,3,4,5,5,1,2
5,4,4,6,5,1,3
5,5,2,6,6,7,4
5,6,1,6,6,10,4
```

Fig. 1. The Weka database after the transformation

The *run information* part contains general information about the scheme used, the number of instances (48) and attributes (7) as well as the attributes names, as presented in figure 2.

```
=== Run information ===

Scheme:       weka.classifiers.trees.SimpleCart -S 1 -M 2.0 -N 5 -C 1.0
Relation:     average_temp
Instances:    48
Attributes:   7
              year
              month
              pres_average
              umidit_relative
              clouds_cant
              precipitations
              average_temp
Test mode:    evaluate on training data
```

Fig. 2. The Run Information output

The second part of the output is represented by the CART decision tree (figure 3).

```
=== Classifier model (full training set) ===

CART Decision Tree

month=(1)|(2)|(3)|(12)|(4)
| pres_average=(5)|(4)|(1)|(2)|(6): 2(13.0/3.0)
| pres_average!=(5)|(4)|(1)|(2)|(6): 3(4.0/0.0)
month!=(1)|(2)|(3)|(12)|(4): 4(23.0/5.0)

Number of Leaf Nodes: 3
Size of the Tree: 5
Time taken to build model: 0.2 seconds
```

Fig. 3. The Classifier model

This tree is interpreted using the If-Then rules:

If (*month* = 1 or 2 or 3 or 12 or 4) and (*pres_average* = 5, 4, 1, 2, 6) then *average_temp* = 2;

If (*month* = 1 or 2 or 3 or 12 or 4) and (*pres_average*! = 5, 4, 1, 2, 6) then *average_temp* = 3;

If (*month*! = 1 or 2 or 3 or 12 or 4) then *average_temp* = 4.

Furthermore, with the coding rules used for the database, these can be explained:

If the *month* is January (1) or February (2) or December (12) or April (4) and the *average pressure* is in one of the next intervals: (1020, 1024) (5), (1015, 1019) (4), (1000, 1004) (1) or (1025, 1030) (6) then *average_temp* is in (15, 19) (2) interval.

The second rule is:

If the *month* is January (1) or February (2) or December (12) or April (4) and the *average pressure* is not (!=) in one of the next intervals: (1020, 1024) (5), (1015, 1019) (4), (1000, 1004) (1) or (1025, 1030) (6) then *average_temp* is in (20, 24) (3) interval.

And the third rule can be explained:

If the *month* is not January (1) or February (2) or December (12) or April (4) then *average_temp* is in (25, 29) (4) interval.

Beside this, Weka provides some complementary information about the percent of correctly as well as incorrectly classified instances. In this example, out of a total of 48 instances, only 40 have been correctly classified meaning 83.3333 %. This summary is presented in figure 4, along with some important statistical parameters. One of it is *Kappa statistic*, a measure of agreement between two individuals, with a 0.7257 value; other parameters are *mean absolute error*- a quantity used to measure how close forecasts or predictions are to the eventual outcomes, *root mean squared error* - a good measure of the model's accuracy*, root relative squared error* - the average of the actual values, *relative absolute error* - similar to the relative squared error.

The forth part of the output, presented in figure 5, contains information regarding the detailed accuracy by class. Here are detailed information concerning the next statistical parameters:

- *TP Rate (True positive rate)* – the report of the positive instances classified as positive. There have been classified as positive 92.8% of the positive instances from class 2, 44.4% from class 3 and 100% from the forth class. The best percentage is for the last class which means that all the positive instances were classified as positive.

- *Precision* – the number of correctly classified instances divided by the whole classified instances number. For example, the precision value is 0.813 for class 2, is 1 for class 3 and 0.821 for class 4.
- *Recall* – the same with TP Rate
- *FP Rate (False Positive Rate)* – the report of the negative classified instances as positive. In our example, for class 2 this report value is 0.088, meaning only 8.85 % from the negative instances have been classified as positive and for class 4 is 0.2.
- *F-Measure* is a measure of a test's accuracy and is determined using the formula:

$$(2 * \text{TP Rate} * \text{Precision}) / (\text{TP Rate} + \text{Precision}) \qquad (1)$$

The values are: for class 2 is 0.929, for the third class is 0.615 and for class 4 is 0.902.

- *ROC Area (Receiver Operating Characteristic Area)* – The ROC curve is given by the TP Rate and FP Rate. The area under the ROC Curve (AUC) is a method of measuring the performance of the ROC curve. If AUC is 1 then the prediction is perfect; if it is 0.5 then the prediction is random. Analyzing our output we conclude that even if the prediction in this case is not perfect, it is not random as well. The best prediction was for class 2 – 0.924 and the "week" prediction is 0.717 for class 5. Between this extremes are the values 0.9 for class 4 and 0.796 for class 3.

```
=== Evaluation on training set ===  == Summary ===

Correctly Classified Instances      40              83.3333 %
Incorrectly Classified Instances     8              16.6667 %
Kappa statistic                     0.7257
Mean absolute error                 0.0813
Root mean squared error             0.2016
Relative absolute error             42.1627 %
Root relative squared error         66.0774 %
Total Number of Instances                 48
```

Fig. 4. Evaluation on training set

If the algorithm can't predict something for classes 1, 6 and 7 the report was completed with question marks.

```
=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area     Class
 0        0        0          0       0           ?            1
 0.929    0.088    0.813      0.929   0.867       0.924        2
 0.444    0        1          0.444   0.615       0.796        3
 1        0.2      0.821      1       0.902       0.9          4
 0        0        0          0       0           0.717        5
 0        0        0          0       0           ?            6
 0        0        0          0       0           ?            7
```

Fig. 5. Detailed accuracy by class

This information can be used in prediction work. Thus, if we know the month and the average pressure value, using this decision tree, we can estimate the future value for the average temperature.

Conclusions

This paper presents an example of a decision tree building process, one of the most common data mining techniques. We have tried to highlight the way the stored data about past events can be used in the forecast of the future ones. We can predict, with a certain accurate, the average temperature for a future month if we have data regarding some important aspects of the weather. For that we can use a decision tree built with CART algorithm implemented in a specialized software in data mining - Weka.

Future work can include the extension of the database with other important weather parameters like wind speed, wind direction or radiation. Beside this aspect, we can enlarge our database with records from other years not only from 2002 to 2005. Having all this improvements in mind, we can increase the precision in building the decision tree and the weather prediction based on it.

References

1. * * * - *Introduction to Data Mining and Knowledge Discovery, Third Edition*, Two Crowds Corporation, http://www.twocrows.com/intro-dm.pdf, accessed on 12 April 2009
2. * * * - *Data mining Models and Algorithms,* http://www.huaat.com/english/datamining/D_App.htm, accessed on 13 April 2009
3. * * * - *Weka Software Documentation*, http://www.cs.waikato.ac.nz/ml/weka/, accessed on 23 April 2009
4. * * * - *Hong Kong Observatory*, http://www.weather.gov.hk/wxinfo/pastwx/extract.htm, accessed on 15 April 2009

Un arbore de decizie pentru predicția stării vremii

Rezumat

Un arbore de decizie este un instrument de suport decizional folosit deseori deoarece este simplu de înțeles şi de interpretat. Arborii de Clasificare şi Regresie – CART – reprezintă o metodă formată dintr-o colecție de reguli bazate pe valorile anumitor variabile din setul de date folosit pentru modelare. În această lucrare este prezentată o aplicație CART în predicția vremii. A fost aleasă colecția de date înregistrate în Hong Kong. Această bază de date conține înregistrări din 2002 până în 2005. Pentru construirea arborelui de decizie a fost folosit un software gratuit dedicat data mining – Weka. Apoi, sunt prezentate atât arborele de decizie cât şi rezultatele statistice referitoare la datele folosite în construirea modelului decizional.