| BULETINUL<br>Universităţii Petrol – Gaze din Ploieşti | Vol. LIX<br>No. 1/2007 | 51 - 56 | Seria<br>Matematică - Informatică - Fizică |
|---|---|---|---|

# A Practical Implementation of a Data Mining Technique

Elia Petre

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

## Abstract

*Over the last years, because the increase of the electronical data was exponential, there was important to be discovered new methods to handle databases. In this paper I have tried to note down the most popular data mining techniques, their importance as well as their characteristics. The practical side consists of a system which implements one of the most well-known algorithms: ID3. Using values of variables about weather conditions, it generates a decision tree which can be used in weather forecasts. The way this algorithm was built and the experimental results are explained here.*

**Key words**: *data mining, decision trees, ID3, JDBC*

## Introduction

One of the most important thing of the past two decades is the huge increase in the amount of electronical information or data. Nowadays, with so many computers used in every human activity, it is estimated that this accumulation of data in the world should double every 20 months.[1] If taking into account the number of computers connected to the Internet we find out that the size and the number of database increase even faster.

Over the past few years, human analysts discover that they can no longer process the large databases without some special tools and they developed Data mining, some techniques that “automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst.”[7]

Its importance was anticipated by a recent Gartner Group Advanced Technology Research Note that listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."[5]

After a short presentation of some data mining techniques, this article presents how one of the most popular data mining algorithms is used in meteorology predictions. Having information about the past weather conditions and about its influence on the decision to play tennis or not, we can find out if we can go on the tennis court or not, using values of some meteorology parameters and the decision tree built with a data mining technique.

## Brief Overview of Data Mining Techniques

There are several techniques that are used in data mining, each one having advantages but also disadvantages. To find out which one is most appropriate for our case, when we want to use our databases in a decision-make process, it is good to know that "data mining doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods". [6] So, having information about our data business and data mining techniques we can decide what we will use. Or we can try them all (if we have enough time, money and data) and find out which one is the best in our case.

The most popular data mining techniques are:

*Artificial neural networks*: A definition is given by Kurt Thearling in his paper *An Introduction to Data Mining: Discovering Hidden Value in your Data Warehouse:* "Non-linear predictive models that learn through training and resemble biological neural networks in structure."[5] (Actual biological neural networks are incomparably more complex.) Neural nets may be used as well in classification problems or for regressions.

*Decision trees*. A decision tree is a tree in which every branch is a choice and every leaf is a decision; this kind of trees is often used for information gain in a decision-make process. There are two kinds of decision trees: the classification trees that find the most appropriate class for each case that is analyzed and regression trees used to predict continuous variables. [6]

A classification problem consists of four main components:

- The categorical output variable or the "dependent" variable; this is what we want to predict;
- The "independent" or "predictor" variables, used to predict the output variable;
- learning dataset, a database which includes values for both the outcome and predictor variables;
- test or future dataset, which consists of another data for whom we would like to be able to make accurate predictions.

For each step, a decision tree chooses a split using a "greedy" algorithm. This means that the split decision ones made at each node, it is never revisited, so each split is dependent on its predecessor. In other words, we can find another final solution every time we use this method if the first split is made differently [6].

*Genetic algorithms*: These techniques are named after the genetic process: the combination of the genetic information in the actual generation in order to find new and better individuals, the mutation that takes place in the genetic code and the natural selection. [5]. It also includes the "chromosome" in which it is placed all the necessary information for building a model (the next generation) until the best is found.

*Multivariate Adaptive Regression Splines (MARS)*. Once CART (one of the most popular decision tree) was released, one of the inventors, Jerome H. Friedman, developed MARS, a method designed to improve CART's shortcomings.

*Rule induction*: Using statistical significance of a variable in database, this method creates if-then rules. Exemple: if wind= strong then play=no.

*Nearest neighbor method and memory-based reasoning (MBR)* A technique that examines some number – the k in k-nearest neighbor – and decide in which class to place a new case. In this method a neighbor is a similar case. After discovering where most of its neighbors are, the new case is assigned to the same class [6].

*Logistic regression* is a generalization of linear regression, when we have to predict binary variables (with values such as 0/1 or yes/no).

*Generalized Additive Models (GAM)* is that models which can extend both linear and logistic regression. In this method the model is a sum of possibly non-linear functions, one for each variable used in the prediction. [6] This class of methods can be used for the classification of a binary output variable as well for its regression.

*Discriminant analysis* In 1936, R. A. Fisher published a method that had been helping him to classify the famous Iris botanical data into three species. He considered classes as hyper planes that are separated by lines in two dimensions, by planes in three etc. It is a simple method and easy to interpret because all you have to do is to determine where a point falls: on which side of the line (or hyper plane). [6]

# Example of Application in Meteorology

## System Description / Architecture

For a small application of data mining in meteorology the DM_ID3 system was built. Its architecture is shown in Fig. 1: a database from which we extract values for both dependent and independent variables using JDBC (Java Database Connectivity) – an interface with the SQL Server, DM_ID3 algorithm – based on a popular method to create a decision tree (ID3 algorithm) and a result knowledge database which can be used in a decision-make process.

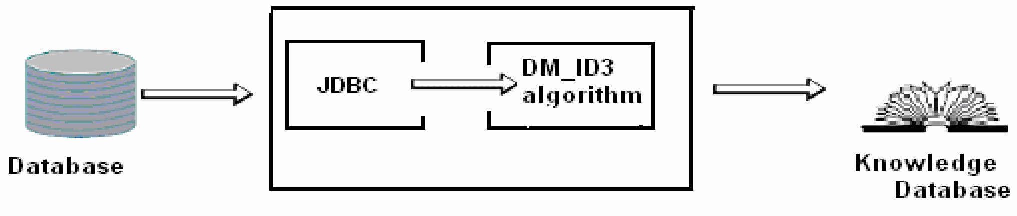


**Fig. 1.** DM_ID3 Architecture

### User Interface

One of the purposes of this system is to integrate it in a web site dedicated to data mining techniques, as an example of how a decision tree is built. Having this in our mind, we have decided that the best choice to implement this system is an applet built in Java Builder 5 that should be connected to a database in Microsoft SQL Server. Knowing that collecting, exploring and selecting the right data are critically important, we have created the database called *Vreme.* The attributes are: *in_general (outlook), temp(temperature), vant(wind), umiditate(humidity), se_joaca(play)* and their possible values: {*insorit(sunny), innorat(overcast), ploua(rain)*}, {*cald(hot), normala(mild), rece(cool)*}, {*slab(week), puternic(strong)*}, {*crescuta(high), normala(normal)*} The last variable depends on the values of the other four variables and it refers to the possibility of playing tennis(DA) or not (NU).

Java Builder 5 is an excellent software which can be used when you want to build a database application, being easy to use and understand. To connect the applet to our database we used a standard interface: JDBC (Java Database Connectivity) which contains a package with classes and methods dedicated to SQL databases: java.sql. In this package there are standard implementations for the connection with SQL server helpful for the developers of databases applications.

Furthermore, we have decided to use the ID3 algorithm to generate a decision tree as a data mining technique.

## The ID3 Algorithm

ID3 Algorithm builds a decision tree by starting a top-down search through the given sets to test each attribute at every tree node. The algorithm uses a greedy search, that is, it chooses the best attribute and never reconsiders the choices he made.

But how does ID3 algorithm decide which attribute is the most useful for classifying a given sets? He uses a metric – information gain. To find out the information gain for each attribute and which argument comes first, we have to determine the entropy of the attributes. The entropy is a numeric value very important for estimating the amount of information gain if at one level is used a certain attribute [8].

The main steps of the ID3 algorithm are:

- For each attribute in the database, compute its entropy;
- The current node is the attribute (A) with highest information gain;
- For every value of the attribute A build a subtree;
  if *A=value1* then generate subtree1;
  if *A=value2* then generate subtree2, etc.
- For each subtree, repeat this process from the first step;
- Every time a new node is created in the tree with a variable, that attribute is removed from the variables group;

The process stops when there are no attributes left [8].

Once we understand the ID3 algorithm, all we had to do is to develop the application.

So the first step was to establish the connection to the SQL server database. This was included in the Start method of the applet. To be sure that this crucial step is successfully implemented we used the bottom-left field of the applet to show the connection or the result of the exception throwing if there are some problems (Fig. 3).

Assuming that the applet was connected to the database, we can see the values of the variables included in our table *Meteo* using the up–right field of the applet (Fig. 3).

Next we created a data reading method which populated dedicated structures with the values of variables extracted from our table.

Then we generated the decision tree by using the entropy and the information gain for each node the tree is split. The result tree is:

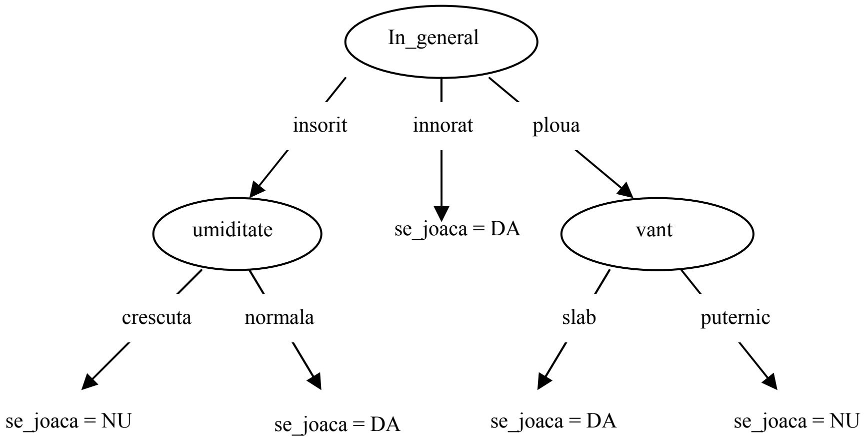


**Fig. 2.** The decision tree

The variable *in_general* was selected first because its information gain was 0.48, the highest. Then for each value of it a new subtree was built:

- the subtree1 with value *'insorit'* for attribute *in_general*;
- the subtree2 with value '*innorat'* for attribute *in_general*;
- the subtree3 with value *'ploua'* for attribute *in_general*;

In the subtree1, the attribute *umiditate* was selected as a decision node and another two subtrees are created: one for the value *crescuta* of the attribute *umiditate* witch leads to the leaf-node '*se_joaca* = NU' and the other one for *umiditate* = *normala* and 'se_joaca = DA'.

The decision tree is easily interpreted, reading from the root to the leaves and using the If-then rules format:

If (*in_general* = *insorit*) and (*umiditate* = *crescuta*) then *se_joaca* = NU;
If (*in_general* = *insorit*) and (*umiditate* = *normala*) then *se_joaca* = DA;

The same way were built the subtree2 and the subtree3.

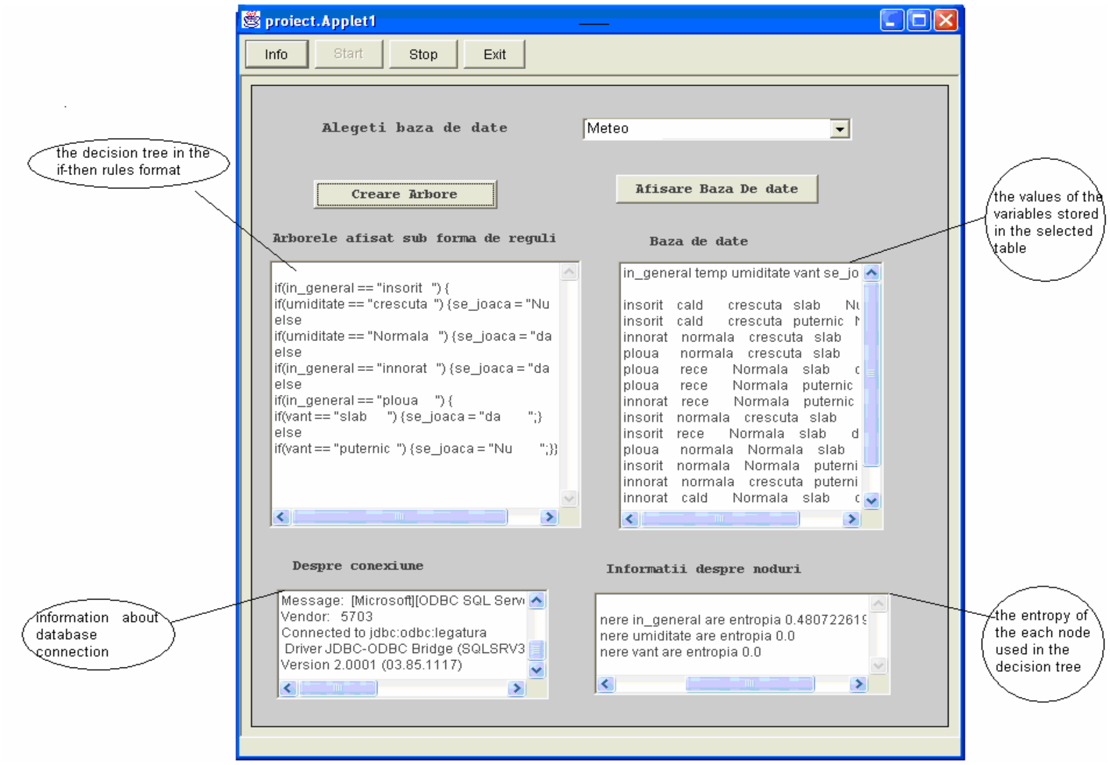
If (*in_general* = *ploua*) and (*vant* = *slab*) then *se_joaca* = DA;
If (*in_general* = *ploua*) and (*vant* = *puternic*) then *se_joaca* = NU;
If (*in_general* = *innorat*) then *se_joaca* = DA;

In our applet the decision tree is displayed using the If-Then rules in the top-left field.

## Experimental Results

The applet showed in Fig. 3, having implemented all the methods, offers to the user the possibility:

- to select a certain table from *Vreme* database;
- to find out if the connection to that table was successfully established;



the decision tree in the if-then rules format

```
if(in_general == "insorit  ") {
if(umiditate == "crescuta  ") {se_joaca = "Nu
else
if(umiditate == "Normala   ") {se_joaca = "da
else
if(in_general == "innorat  ") {se_joaca = "da
else
if(in_general == "ploua    ") {
if(vant == "slab     ") {se_joaca = "da      ";}
else
if(vant == "puternic  ") {se_joaca = "Nu      ";}}
```

the values of the variables stored in the selected table

```
in_general temp umiditate vant se_jo

insorit   cald     crescuta  slab     Nu
insorit   cald     crescuta  puternic N
innorat   normala  crescuta  slab
ploua     normala  crescuta  slab
ploua     rece     Normala   slab     c
ploua     rece     Normala   puternic
innorat   rece     Normala   puternic
insorit   normala  crescuta  slab
insorit   rece     Normala   slab     d
ploua     normala  Normala   slab
insorit   normala  Normala   puterni
innorat   normala  crescuta  puterni
innorat   cald     Normala   slab     c
```

information about database connection

```
Message:  [Microsoft][ODBC SQL Serv
Vendor:   5703
Connected to jdbc:odbc:legatura
 Driver JDBC-ODBC Bridge (SQLSRV3
Version 2.0001 (03.85.1117)
```

the entropy of the each node used in the decision tree

```
nere in_general are entropia 0.48072261
nere umiditate are entropia 0.0
nere vant are entropia 0.0
```

**Fig. 3** Experimental results

- to display the values of the variables stored in the selected table;
- to observe the way that the decision tree is built: at every split it is showed the variable applied and its entropy;
- to see the If-Then rules generated.

Having the decision tree, the user can easily decide if he can play tennis or not.

For example, if he notices that outside is sunny (*in_general = insorit*) and the wind is week (*vant = slab*) then he can go to the tennis court.

## Conclusion and Future Work

Our DM_ID3 system is a simple example of a data mining technique meant to help to a better understanding of a decision tree building process. This way we try to give you an idea about the way the stored data about past events should be used for the prediction of the future ones. Knowing the values of the most important aspects of the weather, using a decision tree we can offer a support for the future decision-make process.

For the future we plan to develop a complex software which includes other data mining techniques able to use the amount of data, to classify and predict them in support of our decisions.

It is estimated that in the future, data mining will become as popular and easy to use as e-mail. Maybe we will be able to use those techniques to find out which plane tickets are better, to track down our high school best fiend's phone number or to achieve the best products.

## References

1. Dilly, R. - *Data mining - An introduction,* Student Notes, Queens University, Belfast, 1995
2. Petre, E. - *Tehnici de tip data mining,* lucrare de licență, Catedra de Informatică, Universitatea Petrol – Gaze Ploiești, 2006
3. Oprea, M. - *Sisteme bazate pe cunoştinţe*, Editura Matrix Rom, Bucureşti, 2002
4. Quinlan, R. - Induction of decision trees, *Machine Learning,* Vol.1*,* 1986
5. Thearling, K. - *An Introduction to Data Mining: Discovering Hidden Value in your Data Warehouse,* www.thearling.com/text/dmwhite/dmwhite.htm, accessed 28 April 2007
6. Two Crows Corporation - *Introduction to Data Mining and Knowledge Discovery*, Third Edition, tutorial booklet, ISBN 1-892095-02-5, 2006
7. * * * - *Data Mining,* http://www.megaputer.com/dm/dm101.php3, accessed 12 May 2007
8. * * * - *The ID3 Algorithm,* http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm, accessed 30 April 2007

# O Implementare Practică a unei Tehnici de Tip Data Mining

## Rezumat

*În ultimii ani, datorită creşterii exponenţiale a datelor stocate în format electronic, necesitatea apariţiei unor noi metode de lucru cu baze de date a devenit tot mai pregnantă. În acest articol am încercat să amintesc cele mai populare metode data mining şi principalele lor caracteristici. Un exemplu practic a fost crearea unui sistem care implementează unul dintre cei mai cunoscuţi algoritmi în literatura de specialitate: ID3. Acesta având date înregistrate despre starea vremii, generează un arbore de decizie care poate fi folosit în predicţiile meteorologice. Modul în care acesta a fost implementat, cât şi rezultatele sale sunt explicate aici.*