BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LIX No. 1/2007	43 - 50	Seria Matematică - Informatică - Fizică

Predicting Infractionality Rate by County Using ID3 Algorithm

Daniela Şchiopu

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

Abstract

This paper presents a classification model which allows us to predict, with certain probability, an infractionality rate specific to a certain Romanian county for which we know population mean of that county, unemployment rate and average net nominal monthly salary earnings. This is made on the basis of statistic data of some counties, from all development regions of the country.

Key words: *ID3 algorithm, decision tree, probability, data mining*

Introduction

This paper considers the following situation: having certain date about a county, such as, the population mean of that county, the unemployment rate and the average net nominal monthly salary earnings, we must get (forecast) an infractionality rate specific to that zone. The problem is one of classification.

For this work, we used statistic data (from 2004) being on the web sites of National Institute of Statistics, relating to 15 counties from the eight development regions of the country. Using the statistic dates for these counties, relating to average net salary on economy from 2004, unemployment rate at country level, population mean on the county and average of infractionality rate, we set three intervals for each variable, for arrange the found dates in just three categories, for calculus easiness.

Using WEKA (Waikato Environment for Knowledge Analysis), realized at the University of Waikato in New Zealand [3], we did an implementation of a data mining technique, starting from the training data, namely ID3 algorithm. After running WEKA application, we obtain a lot of statistic results, including the decision tree; after this, we can get decision rules for classification of the next dates.

After that, giving statistic dates about a certain county, namely just population mean, average net nominal monthly salary earnings and the unemployment rate, we can determine the infractionality rate for that zone.

The paper intends to be an application with good results referring to determine unknown values (in this case, the infractionality rate), knowing all the other dates (population mean, unemployment rate and average net nominal monthly salary earnings) for a certain case. This is

supported by the importance of the used technique, one of the simplest learning algorithms, because it is modelled on base of human cognition, namely ID3 algorithm.

This paper is based on the obtained results after the dissertation elaboration, for finalize master thesis of the author.

Decision Trees – Data Mining Technique

Data mining is a process that uses a variety of information analyses for discover models and relations between these, which can be used for some valid predictions.

The decision trees are arborescent structures that represent decision sets. These decisions generate rules for the classification of data sets. The main decision methods include: *Classification And Regression Trees* (CART), *Chi Square Automatic Interaction Detection* (CHAID), *C4.5* . CHAID can produce a tree with multiple sub-nodes for each division. CART requires more little preparation of data than CHAID, but it always divides data set into just two parts. ID3 algorithm is a predecessor of C4.5 . C4.5 comes from intelligent machines universe, capable of learning and it is based on information theory. [1]

A decision tree is a tree in which each branch represents a choice variant, and each leaf node is a decision. This kind of trees is often used with the object of information gain in a taking decision process. [2]

An algorithm based on decision tree divides the data set with the object of build up a model that classifies each record in terms of target field or variable.

Learning about decision trees is a method of discreet value approximate, in which the learning function is a decision tree. This method is one of the most used techniques of inductive inference. [5]

The decision trees classify the instances through their crossing from the root to the leaf nodes. It starts from the root, testing its attribute, and then shifting to the tree branches, according to the values of the attribute in the given data set.

ID3 Algorithm

ID3 is an algorithm of inductive build up decision trees. The basic idea in ID3 is selecting the most important attribute the first time. The most important, meaning that which differentiates examples the most is not yet classified. The measure of chosen importance on this algorithm is entropy, which is calculated for each subset.

As a consequence, ID3 chooses the attribute considered the most important and divides the examples in subsets corresponding to the possible values of the attribute. Next, it uses ID3 recursively to obtain subset and list remaining attributes. The recursive proceeding ends when all the examples from one subset have the same classification. If all attributes end and the subset still contains examples of different classification, this means there are examples with the same description, but with different classification. This thing can be caused by many reasons:

- few data are incorrect;
- data are correct, but the attributes are insufficient;
- data are correct, but the classification implies a certain level of indeterminism.

A mathematical definition of ID3 tells that it algorithmically determines, the greatest gain in information content, while reducing system entropy. The concepts of information content and entropy are concepts that often appear in many aspects of computer science, including information theory, algorithms and data compression. Shortly, information content refers to the

information quantity (data quantity), which is contained by each unit of representation (usually, this unit is the bit) and entropy is the minimum number of representation units, required for communicate a certain data set.

The entropy is measured on a scale from 0 to 1; it can be defined mathematically, too. Let m be the number of elements in W and n_a be the number of instances of the element a in W. Therefore, the probability p_a of choosing an element a from W is given by this:

$$p_a = \frac{n_a}{m}$$

For a simple system with subclasses c_i , i = 1,2,…,C, the entropy of system can be defined as:

$$Entropy = \sum_{i=1}^{C} - p_i \log_2 p_i \quad (1)$$

For example, in a data set with two distinct values, let be M, where there is an equal probability of encountering each of the values, the entropy is calculated as follows:

$$Entropy(M) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1 \quad (2)$$

At each step, we must choose the attribute that leads to the smallest entropy.

Predicting the Infractionality Rate on the County

Let us consider the follow situation. Giving certain data about a county, such as, *population mean*, *unemployment rate* and *average net nominal monthly salary earnings*, we can determine an infractionality rate of that zone.

For this, we used statistic data present on the web sites of the National Institute of Statistics, from 2004, relating to a few counties from the eight development regions of the country, namely, from North-East region (Iaşi and Neamț counties), South-East (Brăila, Vrancea and Tulcea counties), South-Muntenia (Prahova and Argeş counties), South-West Oltenia (Vâlcea and Mehedinți counties), West (Timiş and Arad counties), North-West (Cluj and Sălaj), Center (Alba county) and Bucureşti-Ilfov (the capital and Ilfov county). The counties have been chosen so that they represent a representative pattern for the entire population (in this case, all counties) and homogeneous, for a better exactitude of the data. The training data are presented in **Table 1**.

After the data taken from the National Institute of Statistics, for a better data operation and for an easiness in calculus and in building the decision tree, we established for each variable, three intervals, in conformity with: population mean (by the counties), medium net salary (in 2004), unemployment rate (at the country level) and the mean of infractionality rate (per each county).

These values are:

- population mean: 510.000 ;
- medium net salary on economy: 582,800 ;
- unemployment rate: 8% ;
- mean of infractionality rate:1069.

Thus, we have the results presented in **Table 2**.

The variable *Population_mean* contains the entire population resident in a region (the population mean from the beginning of the year and from the end of the year, multiplied with 1000).

Table 1. Training Data

Development region	County	Population mean	Average net nominal monthly salary earnings (RON/employee)	Unemploy-ment rate (%)	Infractionality rate (100 000 people)
North-East	Iaşi	821.621	571,200	7,1	987
North-East	Neamț	570.367	502,883	7,2	1279
South-East	Brăila	371.749	533,634	8,7	806
South-East	Vrancea	394.286	528,914	4,2	1193
South-East	Tulcea	713.825	624,997	5,9	1331
South-Muntenia	Prahova	829.026	602,967	6,6	791
South-Muntenia	Argeş	647.437	601,944	6,8	825
South-West Oltenia	Vâlcea	416.908	547,725	7,6	585
South-West Oltenia	Mehedinți	305.901	647,948	10,2	1184
West	Timiş	661.593	609,023	2,6	796
West	Arad	460.466	538,350	3,6	1289
North-West	Cluj	686.825	620,832	5,1	1104
North-West	Sălaj	247.796	568,646	6,2	955
Center	Alba	382.971	517,510	10	1184
Bucureşti-Ilfov	Bucureşti and Ilfov county	2.204.996	714,167	3,4	954

Table 2. The intervals for the four variables

	Small	Medium	Large
Population mean	(0..380.000)	[380.000 .. 510.000]	>510.000
Average net nominal monthly salary earnings	(0..540,000)	[540,000 .. 582,800]	>582,800
Unemployment rate	(0..5)	[5..8]	>8
Infractionality rate	(0..850)	[850..1069]	>1069

Framing values for each variable, considering **Table 2**, initial data will be transformed into follows data, as we see in **Table 3**.

Next, we encoded the values „small", „medium", „large", with 0, 1 and 2. We used these modified data when building of ARFF file, used in WEKA.

Then, having all necessary data, we will try to build up the proper decision tree, using ID3, enounced in the previous chapter. For this, it must determine the entropies for each subset from the training data. For a better data operation, we will transform the real data into nominal data, conformity with certain intervals. This change will be useful to determine ID3 tree, too, with WEKA.

Table 3. Training Data, modified as they take part in specified intervals

Development region	The county	Population mean	Average net nominal monthly salary earnings (RON/employee)	Unemploy-ment rate (%)	Infractionality rate (100 000 people)
North-East	Iaşi	large	medium	medium	medium
North-East	Neamț	large	small	medium	large
South-East	Brăila	small	small	large	small
South-East	Vrancea	medium	small	small	large
South-East	Tulcea	small	medium	medium	large
South-Muntenia	Prahova	large	large	medium	small
South-Muntenia	Argeş	large	large	large	small
South-West Oltenia	Vâlcea	medium	medium	medium	small
South-West Oltenia	Mehedinți	small	large	large	large
West	Timiş	large	large	small	small
West	Arad	medium	small	small	large
North-West	Cluj	large	large	medium	large
North-West	Sălaj	small	medium	medium	medium
Center	Alba	medium	small	large	large
Bucureşti-Ilfov	Bucureşti and jud. Ilfov	large	large	small	medium

Results Obtained With WEKA

WEKA (*Waikato Environment for Knowledge Analysis*) is an open source program, written in Java, developed by the University of Waikato in New Zealand, distributed under GNU - General Public License. [3] WEKA is a software tool that implements data mining algorithms. It contains tools for classification, regression, clustering, association rules, data visualisation. It comes with a *Graphical User Interface* (GUI), but it can be used Java code, too.

WEKA works with ARFF files (*Attribute Relation File Format*). An ARFF is a text file that describes a list of instances that divide a set of attributes.

In our case, using notation convention, ARFF file can contain the follows:

@relation infractionality

@attribute population_mean {0,1,2}

@attribute salary_earnings {0,1,2}

@attribute unemployment_rate {0,1,2}

@attribute infractionality_rate {0,1,2}

@data 1,1,1,0
2,1,1,1 0,2,2,2
2,0,1,2 2,2,0,0
0,0,2,0 1,0,0,2
1,0,0,2 2,2,1,2
0,1,1,2 0,1,1,1
2,2,1,0 1,0,2,2
2,2,2,0 2,2,0,1

We can observe that the last attribute, *infractionality_rate*, is actually the class that is to be predicted, namely dependent variable (the target variable). The data from section @data are those obtained in Table 3 and 4.

Loading the file ARFF in Weka 3.5.5, with the four attributes, and choosing the ID3 method, the classifier model which results, contains the follows: the decision tree, a summary of evaluation on training set (Kappa statistic and the errors), detailed accuracy by class (TP rate, FP rate, precision, recall, F-measure, ROC area) and the confusion matrix.

The decision tree supplied by WEKA is:

salary_earnings = 0
| population_mean = 0: 0
| population_mean = 1: 2
| population_mean = 2: 2
salary_earnings = 1
| population_mean = 0: 1
| population_mean = 1: 0
| population_mean = 2: 1
salary_earnings = 2
| unemployment_rate = 0: 0
| unemployment_rate = 1: 0
| unemployment_rate = 2
| | population_mean = 0: 2
| | population_mean = 1: null
| | population_mean = 2: 0

Building the tree structure after these results, we obtained:

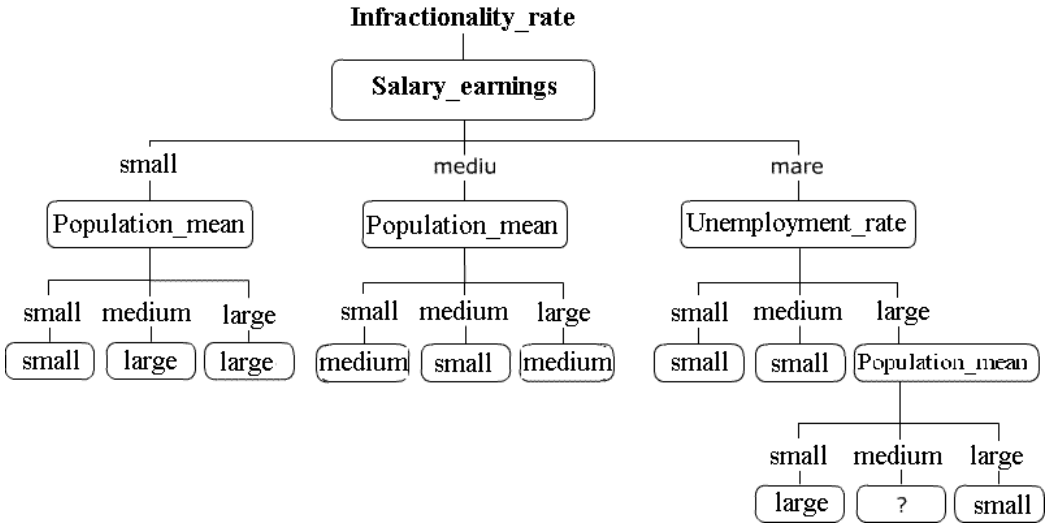

Fig. 1. ID3 Decision Tree

As we can see in the decision tree obtained in Weka (Fig. 1), it results a set of decision rules:

IF *Salary_earnings*<540 AND *Population_mean*<380.000 THEN

Infractionality_rate<850

IF *Salary_earnings*<540 AND *Population_mean* ∈ [380.000, 510.000] THEN

Infractionality_rate>1069

IF *Salary_earnings*<540 AND *Population_mean*>510.000 THEN

Infractionality_rate>1069

IF *Salary_earnings* ∈ [540, 582,8] AND *Population_mean*<380.000 THEN

Infractionality_rate ∈ [850, 1069]

IF *Salary_earnings* ∈ [540, 582,8] AND *Population_mean* ∈ [380.000, 510.000] THEN *Infractionality_rate*<850

IF *Salary_earnings* ∈ [540, 582,8] AND *Population_mean*>510.000 THEN

Infractionality_rate ∈ [850, 1069]

IF *Salary_earnings*>510.000 AND *Unemployment_rate*<5

THEN *Infractionality_rate*<850

IF *Salary_earnings*>510.000 AND *Unemployment_rate*∈ [5, 8]

THEN *Infractionality_rate*<850

IF *Salary_earnings*>510.000 AND *Unemployment_rate*>8 AND *Population_mean*<380.000

THEN *Infractionality_rate*>1069

IF *Salary_earnings*>510.000 AND *Unemployment_rate*>8 AND

Population_mean∈ [380.000, 510.000]

THEN „We can't tell anything about Infractionality_rate"

IF *Salary_earnings*>510.000 AND *Unemployment_rate*>8 AND *Population_mean*>510.000

THEN *Infractionality_rate*<850

If we take for Mureş county, in 2005, data about salary earnings, population mean and unemployment rate [4], we would observe that we don't know infractionality rate for this region.

Table 4. The Three Attributes for Mureş in 2005

County	Average net nominal monthly salary earnings		Population mean		Unemployment rate	
Mureş	671	large	583.383	large	4,6	small

According to decision tree from *Fig. 1*, it results that prevision for infractionality rate for Mureş, for 2005, is „small", namely is smaller than 850, at 100 000 people.

Conclusions

The present application can forecast an unknown value, on the base of some real, known values, using a pretty simple technique and that doesn't use too many scientific details. We only calculate entropy and information gain for each case, in order to build the decision tree more efficiently. Using WEKA, we have obtained the good results for this problem. As it results from confusion matrix too, the algorithm has classified correctly 80 % of instances (meaning 12) and incorrectly 20 % (meaning 3). The Kappa statistics is 0,6918 (and we know that, in order to consider it important, its value must be higher than 0,6).

For this data mining technique, as future developments, we can mention any application that necessitates prediction of some unknown data using the others known, but not only. In this case, we specify the application of the decision trees together with other areas of interest, such as semantic web.

References

1. * * * - *Building Classification Models: ID3 and C4.5,* http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html, *accessed* 23 May 2007
2. * * * - *Arbori de decizie,* http://eureka.cs.tuiasi.ro/~fleon/BVIA/Arbori%20de%20decizie.pdf, accessed 20 June 2007
3. * * * - Weka, http://www.cs.waikato.ac.nz/~ml/weka/, accessed 12 June 2007
4. * * * - www.galati.insse.ro, accessed 15 June 2007
5. Witten, I.H., Frank, E. - *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, Morgan Kaufmann, 1999

Predicția Ratei Infracționalității pe Județe Folosind Algoritmul ID3

Rezumat

În acest articol construim un model de clasificare; modelul ne permite să prevedem, cu o anumită probabilitate, o rată a infracționalității specifice unei anumite zone din România, pentru care se cunosc media populației din acel județ, rata şomajului şi câştigul salarial nominal mediu net lunar. Predicția se bazează pe date statistice referitoare la câteva județe, din toate regiunile de dezvoltare ale țării.