BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LIX No. 2/2007	37 - 42	Seria Matematică - Informatică - Fizică

The Divorce Rate Prediction using Data Mining Techniques

Mădălina Cărbureanu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: mcarbureanu04@yahoo.com

Abstract

The divorce phenomenon has social and economic implications and that is why is useful it's analysis. The attributes witch may influence this phenomenon can be: the number of marriages, the net medium nominal monthly income, the unemployment rate, the medium age at marriage and the education level index. In this paper an application of data mining techniques is presented so as to highlight the opportunity of using these methods in the field of demography and social statistics, with the final goal of predicting the divorce rate for a certain year, at district level.

Key words: *data mining, decision rules, divorce rate prediction, ID3 algorithm*

Introduction

The data mining techniques are successfully applied to solve concrete problems in real life, problems from various domains, such as: the banking financial domain, the retail trade domain, the health domain, the telecommunications domain, and last but not least, the demography and social statistics domain[1].

In this paper we present an application from the demography and social statistics domain, for predicting the married population general divorce rate for 2005, at district level, rate which is unknown. In order to solve this application, we used the data mining techniques, namely, the ID3 algorithm, which is an algorithm from the decision tree category. To implement this algorithm, we used the Weka (Waikato Environment Knowledge Analysis) software, which is a collection of machine learning algorithms for data mining .

A Few Words about the Divorce Phenomenon

The divorce phenomenon characterises the divorces throng which took place within a population during a certain period of time, usually in a certain year.

In the field of demography, in order to measure divorce intensity, two categories of rates can be calculated, some compared to the entire population, and others which only take into account the population exposed to the divorce risk, therefore the married population.

In order to measure the divorce intensity, the most used indicators are:

- The gross divorce rate, relative measure which is represented by the ratio between the number of divorces and the entire population; the disadvantage is that the denominator doesn't represent the population exposed to the divorce risk, but the entire population;
- The divorce general rate of married population which is the ratio between the number of divorces and the married population exposed to the divorce risk;
- The divorce specific rates which consist of measuring the divorce intensity on communities, ordered by age and sex;
- The ratio between the number of divorces and the number of marriages which took place in a certain year [2].

Because the population exposed to the divorce risk is only the married population, in the case of our application we used the divorce general rate of married population.

Application Example

The application consists in achieving the prediction of the divorce general rate level of married population for 2005, for any district of Romania, using the data mining techniques. For this an algorithm from the decision tree category was used, named ID3. The ID3 algorithm is an inductive technique of Artificial Intelligence for generating decision trees.

The attributes used for the current application are the following: the number of marriages, the net medium nominal monthly income, the unemployment rate, the medium age at marriage, the education level index, and, the target attribute, the married population divorce general rate.

The training data presented in figure 1, were supplied by the National Statistics Institute [3].

Development region	District	Marriages	Net medium nominal monthly income (RON/ Wage earner)	Unemployment rate (%)	Medium age at marriage (years)	Education index	Married population divorce general rate (at 1000 marriages)
NORD-EST	Iasi	5780	571,200	7.1	27.7	0.936	153.8
NORD-EST	Neamt	3712	502,883	7.2	27.05	0.844	311.4
NORD-EST	Botosani	2672	494,313	8.0	26.55	0.839	264.6
SUD-EST	Constanta	6022	624,997	5.9	29.35	0.910	169.9
SUD-EST	Buzau	3228	526,370	6.7	28.25	0.845	243.2
SUD-EST	Tulcea	1676	564,100	5.6	28.85	0.825	251.8
SUD-MUNTENIA	Prahova	4987	602,967	6.6	28.55	0.866	311.4
SUD-MUNTENIA	Dambovita	3458	587,787	6.7	28	0.854	279.9
SUD-MUNTENIA	Giurgiu	1418	540,655	7.3	28.35	0.792	122
SUD-VEST	Dolj	3942	597,262	5.8	28.2	0.902	100.7
SUD-VEST	Olt	2670	583,552	7.7	27.75	0.839	159.2
SUD-VEST	Mehedinti	1875	647,948	10.2	28.7	0.864	273.1
VEST	Timis	4999	609,023	2.6	30.1	0.954	232.8
VEST	Caras-Severin	2844	506,623	9.0	30.35	0.861	207.8
VEST	Arad	3046	538,350	3.6	29.5	0.900	276.4
NORD-VEST	Cluj	4669	620,832	5.1	29.2	0.969	271.4
NORD-VEST	Salaj	1437	568,646	6.2	26.95	0.848	188.6
NORD-VEST	Maramures	3744	500,107	4.6	27	0.853	272.4
CENTRU	Brasov	3916	554,315	10.7	29	0.911	219.1
CENTRU	Covasna	1299	501,616	8.1	28.2	0.838	307.9
CENTRU	Sibiu	2870	579,925	6.3	28.25	0.917	289.9
BUCURESTI-ILFOV	Municipiul Bucuresti	15742	740,465	2.6	30.8	1.058	192.7

Fig. 1. The training data

Because we choose for solving the proposed application the ID3 algorithm, we must build some intervals for each attribute. The mentioned intervals were established taking into consideration the characteristic features of each attribute for the year 2004[3]. Therefore, the average of marriages for each district was 3412, the net medium nominal income per overhaul economy for

2004 was 598,639 RON, the unemployment rate was 8.0%, the average age at the marriage was 28.55, the education level index at the national level for 2004 was 0.889, and the married population divorce general rate at one thousand marriages was 245.8. Knowing the average for each attribute, we build the intervals like this: the values which are bigger than this average belongs to the *big* interval, for establish the *medium* interval we made an approximation of it using the medium values of the training data and the values witch are smaller than these medium values belongs to the *small* interval.

Therefore, the intervals for each attribute are those presented in figure 2.

	Small	Medium	Big
Marriages	(0; 1680)	[1680; 3412]	>3412
Net medium nominal monthly income (RON/Wage earner)	(0; 540,000)	[540,000; 598,639]	>598,639
Unemployment rate (%)	(0; 5)	[5; 8]	>8
Medium age at marriage (years)	<18	[18; 28.55]	>28.55
Education index	(0; 0.844)	[0.844; 0.889]	>0.889
Married population divorce general rate (at 1000 marriages)	(0; 170)	[170; 245.8]	>245.8

Fig. 2. The intervals for each attribute

Because we are working with the Weka software, we must build a file source named *divorce.arff*. The source file must have an *arff* or *xrff* extension, because these are the only extensions accepted by the W eka software. The source file has two distinct sections, the header section and the data section. The header section contains the name of the relation, a list of attributes and their types. The data section contains the values for each attribute. The source file obtained is presented in Fig. 3.

Using the source file *divorce.arff* and applying the ID3 algorithm on the data from this file, we obtain a decision tree, presented Fig. 4.

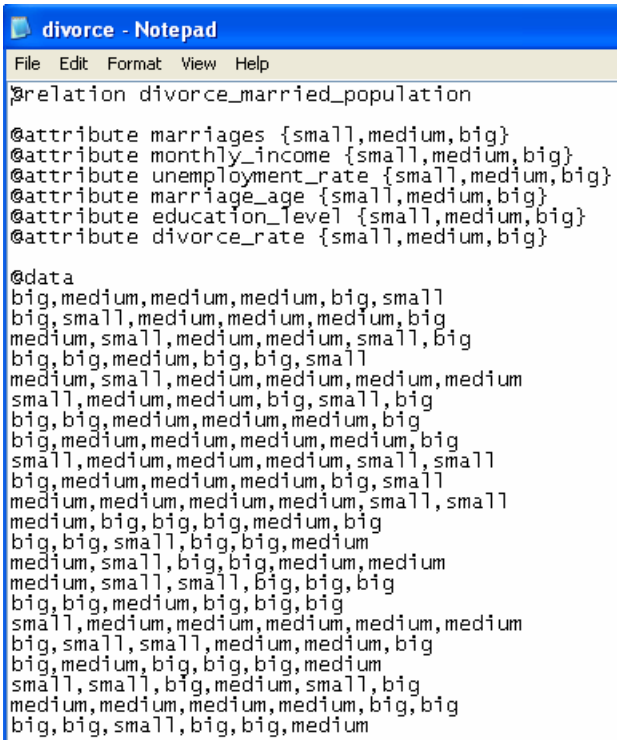

Fig. 3. The source file *divorce.arff*

```
education_level = small
|  monthly_income = small: big
|  monthly_income = medium
|  |  marriage_age = small: null
|  |  marriage_age = medium: small
|  |  marriage_age = big: big
|  monthly_income = big: null
education_level = medium
|  marriages = small: medium
|  marriages = medium
|  |  monthly_income = small: medium
|  |  monthly_income = medium: null
|  |  monthly_income = big: big
|  marriages = big: big
education_level = big
|  unemployment_rate = small
|  |  marriages = small: null
|  |  marriages = medium: big
|  |  marriages = big: medium
|  unemployment_rate = medium
|  |  marriages = small: null
|  |  marriages = medium: big
|  |  marriages = big
|  |  |  monthly_income = small: null
|  |  |  monthly_income = medium: small
|  |  |  monthly_income = big: small
|  unemployment_rate = big: medium
```

Fig. 4. The decision tree

The ID3 algorithm uses a special measure, named *information gain.* The information gain is used for the selection of the best classification attribute. The formula for calculating the information gain is the following:

$$I_E(i) = -\sum_{j=1}^{m} f(i, j) * \log_2 [f(i, j)].$$

The decision tree root knot is *education_level* attribute, because this attribute has the highest information gain, which is 0.285. The information gain is used to pick out the testing attribute, based on the information maximise, or, equivalent, the maximum reduction of entropy at the respective knot. The information gain for the others attributes, is: for *marriages* attribute is 0.019, for *monthly_income* is 0.202, for *unemployment_rate* is 0.225 and for *marriage_age* is 0.092, therefore, the *education_level* attribute is the decision tree root knot.

Using the decision tree presented in figure 4, we can detach a batch of decision rules. To detach these decision rules we use the intervals presented in figure 2. Some of the decision rules are presented in figure 5.

IF education_level ∈ (0 ; 0.844) AND monthly_income ∈ (0; 540,000) THEN divorce_rate>245.8;
IF education_level ∈ (0 ; 0.844) AND monthly_income ∈ [540,000; 598,639] AND marriage_age<18 THEN We can say anything about divorce_rate;
IF education_level ∈ (0 ; 0.844) AND monthly_income ∈ [540,000; 598,639] AND marriage_age ∈ [18 ; 28.55] THEN divorce_rate ∈ (0 ; 170);
IF education_level ∈ (0 ; 0.844) AND monthly_income ∈ [540,000; 598,639] AND marriage_age>28.55 THEN divorce_rate>245.8;
IF education_level ∈ (0 ; 0.844) AND monthly_income>598,639 THEN We can say anything about divorce_rate;
IF education_level ∈ [0.844 ; 0.889] AND marriages ∈ (0 ;1680) THEN divorce_rate ∈ [170 ; 245.8];
IF education_level ∈ [0.844 ; 0.889] AND marriages ∈ [1680 ; 3412] AND monthly_income ∈ (0; 540,000) THEN divorce_rate ∈ [170 ; 245.8];
IF education_level ∈ [0.844 ; 0.889] AND marriages ∈ [1680 ; 3412] AND monthly_income ∈ [540,000;598,639] THEN We can say anything about divorce_rate;
IF education_level ∈ [0.844 ; 0.889] AND marriages ∈ [1680 ; 3412] AND monthly_income>598,639 THEN divorce_rate>245.8;
IF education_level ∈ [0.844 ; 0.889] AND marriages>3412 THEN divorce_rate>245.8;

Fig. 5. Some of the decision rules

Knowing the number of marriages, the net medium nominal monthly income (RON/Wage earner), the unemployment rate (%), the medium age at marriage (years), the education level index (the arithmetic mean between the adult population literacy rate and the gross enrolment ratio for all education levels), we can establish for each district, in which interval from those three presented in figure 2, the divorce rate belongs to. For the decision tree validation, we choose at random the next districts which are different of the districts used like training data: Suceava, Hunedoara, Bihor, Mureş, Bistriţa-Năsăud, Vrancea, Bacău, Galaţi, Călăraşi and Ialomiţa. The statistical data for these districts, for year 2004, are those presented in figure 6 [3].

Using the decision rules presented in figure 5, we have that:

- Suceava, Hunedoara, Mureş, Vrancea, Bacău, Galaţi, Călăraşi and Ialomiţa will have a *big* divorce rate, therefore the *divorce_rate*>245.8;
- Bihor and Bistriţa-Năsăud will have a *medium* divorce rate, therefore the *divorce_rate* ∈ [170; 245.8].

District	Marriages	Net medium nominal monthly income(RON/ Wage earner)	Unemployment_ rate (%)	Medium age at marriage (years)	Education Index
Suceava	5318(big)	526,948(small)	7.8(medium)	27.35(medium)	0.866(medium)
Hunedoara	3104(medium)	670,064(big)	10.8(big)	30.1(big)	0.886(medium)
Bihor	3782(big)	516,642(small)	2.1(small)	28.2(medium)	0.910(big)
Mures	3445(big)	563,449(medium)	4.4(small)	29.15(big)	0.867(medium)
Bistrita-Nasaud	2196(medium)	523,896(small)	6.4(medium)	27.05(medium)	0.846(medium)
Vrancea	2463(medium)	528,914(small)	4.2(small)	28.35(medium)	0.827(small)
Bacau	4902(big)	586,528(medium)	7.0(medium)	28.05(medium)	0.852(medium)
Galati	3973(big)	620,459(big)	9.6(big)	28(medium)	0.876(medium)
Calarasi	1869(medium)	476,578(small)	8.8(big)	28.25(medium)	0.816(small)
Ialomita	1459(small)	546,934(medium)	10.4(big)	28.60(big)	0.835(small)

Fig. 6. The validation data for the selected districts

Using the values of married population divorce general rate at one thousand marriages, values offered by the National Institute of Statistics [3], we have that:

- For Suceava district, the *divorce_rate* value was 253.5, therefore the *divorce_rate*>245.8;
- For Hunedoara, the *divorce_rate* value was 443.6, therefore the *divorce_rate*>245.8;

For Bihor, the *divorce_rate* value was 175, therefore the *divorce_rate* $\in$ [170; 245.8];

- For Mureş, the *divorce_rate* value was 247.9, therefore the *divorce_rate*>245.8;
- For Bistrița-Năsăud, the *divorce_rate* value was 219, therefore the *divorce_rate* $\in$ [170; 245.8];
- For Vrancea, the *divorce_rate* value was 265.1, therefore the *divorce_rate*>245.8;
- For Bacău, the *divorce_rate* value was 374.7, therefore the *divorce_rate*>245.8;
- For Galați, the *divorce_*rate value was 342.3, therefore the *divorce_rate*>245.8;
- For Călăraşi, the *divorce_rate* value was 338.7, therefore the *divorce_rate*>245.8;
- For Ialomița, the *divorce_rate* value was 320.8, therefore the *divorce_rate*>245.8.

We notice that for these districts, the values of the married population divorce general rate obtained using the ID3 algorithm, respectively, the decision rules supplied by this algorithm, correspond to the values offered by the National Institute of Statistics[3].

If we consider, for five random selected districts, for the year 2005, the data offered by the National Institute of Statistics [3], regarding the number of marriages, the net medium nominal monthly income, the unemployment rate, the medium age at marriage and the education level index, we can predict the married population divorce general rate, which is unknown, for each district. The values for each attribute for the districts selected for making the prediction, are those offered by the National Institute of Statistics, being presented in figure 7.

District	Marriages	Net medium nominal monthly income(RON/ Wage earner)	Unemployment_ rate (%)	Medium age at marriage (years)	Education Index
Suceava	5015(big)	639,000(big)	6.0(medium)	27.35(medium)	0.869(medium)
Hunedoara	2797(medium)	761,000(big)	9.4(big)	30.1(big)	0.887(medium)
Bihor	3808(big)	629,000(big)	2.7(small)	28.2(medium)	0.910(big)
Bacau	4906(big)	718,000(big)	6.3(medium)	28.05(medium)	0.855(medium)
Galati	4072(big)	735,000(big)	8.3(big)	28(medium)	0.878(medium)

Fig. 7. The values for each attribute

According to the decision rules presented in figure 5, the prediction of the married population divorce general rate, for the districts presented in figure 7, for year 2005, is the following:

- For Suceava district, the divorce rate prediction, shows that for this district the *divorce_rate* is *big*, therefore the *divorce_rate*>245.8;
- For Hunedoara district, the divorce rate prediction, shows that for this district the *divorce_rate* is *big,* therefore the *divorce_rate*>245.8;
- For Bihor district, the divorce rate prediction, shows that for this district the *divorce_rate* is *medium*, therefore the *divorce_rate* ∈ [170; 245.8];
- For Bacău, the divorce rate prediction, shows that for this district the *divorce_rate* is *big*, therefore the *divorce_rate*>245.8;
- For Galați, the divorce rate prediction, shows that for this district the *divorce_rate* is *big*, therefore the *divorce_rate*>245.8.

In our application the final goal was to predict an unknown rate, namely the married population divorce general rate, which we called *divorce_rate*, for year 2005, for a certain district. Using data mining techniques, we can make predictions for unknown rates with very good results.

Conclusion

Knowing for a certain district the divorce rate for a certain year and using also the divorce rate evolution for the precedent years, we can detach the trends of this rate. Knowing these information's can be taken measures for stopping the growing evolution of this phenomenon. This article shows that the divorce rate value for a district can be predicted also with data mining techniques.

References

1. Gorunescu, F. – *Data mining. Concepte, modele şi tehnici*, Editura Albastră, Cluj-Napoca, 2006
2. Ţarcă, M. – *Demografie*, Editura Economică, Bucureşti, 1997
3. * * * - Institutul Național de Statistică, http://www.insse.ro, accessed 25 August 2007

Predicția Ratei Divorțialității folosind Tehnici de Data Mining

Rezumat

Fenomenul divorțialității are implicații economice şi sociale, motiv pentru care este utilă analiza acestuia. Atributele care pot influența acest fenomen pot fi: numărul casătoriilor, venitul nominal mediu net lunar, rata şomajului, vârsta medie la casătorie şi indicele nivelului de educație. În această lucrare este prezentată o aplicație a tehnicilor de data mining pentru a sublinia oportunitatea utilizării acestor metode în domeniul demografiei şi statisticii sociale, cu scopul final de a prezice rata divortialității pentru un anumit an, la nivel de județ