Leszek MISZTAL
WEST POMERANIAN UNIVERSITY OF TECHNOLOGY, DEPARTMENT OF INFORMATION SYSTEMS ENGINEERING

Rough Set Application for the Tax Payer Classification Rules

Mgr inż. Leszek MISZTAL

Received the M.Sc. degree in Computer Science from Technical University of Szczecin in 1999. Nowadays PhD student at West Pomeranian University of Technology. His current research interests are focused on information theory, data mining mainly in the domain of taxation problems and also issues related with integration of information systems.

e-mail: lmisztal@wi.ps.pl

Abstract

Classification of the tasks for real-world problems becomes possible because of creation and use of more efficient IT systems. It also targets rough set methods as well described with solid mathematical basis for classification tasks. In the presented paper the application of rough set theory with the usage of significance of attributes and decision rule sets for classification of taxpayers is described. There are taken into account the negative or positive results of taxation control, and specific features describing payers are considered. Appropriate choice of data, building the model and its application leads to the specified goal reaching, with better accuracy in comparison to "intuitive" choice. Simultaneously it becomes possible to extract decision rules in the linguistic form, what gives opportunity for easier interpretation of obtained results. As a result of the solution application the more accurate selection of tax payers is obtained. This is of significant meaning for the tax authorities, and this leads for the better observance of the tax law.

Keywords: rough sets, data mining, classification, rules extraction, decision rules.

Zastosowanie teorii zbiorów przybliżonych w zadaniu klasyfikacji podatników

Streszczenie

Rozwiązywanie zadań klasyfikacji dla rzeczywistych problemów stało się możliwe dzięki rozwojowi wydajniejszych systemów informatycznych. Dotyczy to również teorii zbiorów przybliżonych dla zadań klasyfikacji. W przedstawionej publikacji zastosowano zbiory przybliżone, które mają ugruntowaną teorię bazującą na rozszerzeniu teorii zbiorów i definiującą dolne oraz górne przybliżenie, oraz wyznaczającą tabelę decyzyjną do klasyfikacji. Metodę użyto do obliczeń istotności atrybutów oraz reguł decyzyjnych opisujących klasyfikację podatników ze względu na pozytywny lub negatywny wynik kontroli, przy uwzględnieniu specyficznych cech ich opisujących. Odpowiedni dobór danych, budowa modelu oraz jego użycie umożliwiło osiągnięcia zadanego celu ze zwiększoną dokładnością w stosunku do „intuicyjnego" wyboru. Wykorzystanie zbiorów przybliżonych, które wyznaczają wyniki końcowe klasyfikacji w postaci zbioru reguł umożliwiło ich ekstrakcję w łatwo interpretowalnej formie lingwistycznej. W publikacji zastosowano autorskie rozwiązanie programowe bazujące na kolekcjach, tablicach oraz obiektach pośrednich, zaimplementowane dla bazy danych Oracle, dzięki któremu zrealizowano zadanie oraz przedstawiono rezultaty. Dzięki uzyskanym wynikom bazującym na modelu opartym na użytej metodzie możliwe staje się dokładniejsze typowanie podatników funkcjonujących w polskim systemie prawnym i mających problemy podatkowe, których należy poddać kontroli. Tym samym zwiększa się skuteczność egzekwowania prawa podatkowego.

Słowa kluczowe: zbiory przybliżone, eksploracja danych, klasyfikacja, ekstrakcja reguł, reguły decyzyjne.

1. Introduction

Classification is a process of training of the model describing different data classes. It is assumed, that these classes are known and they are determined before classification start. Because of described facts this kind of training tasks is called supervised learning [12]. After creation of the model on the basis of training data it is usable for the production data, where unclassified data have to be assigned to correct classes.

Rough set theory for classification uses notion of approximation space, and lower and upper approximations of set [3]. The approximation space of a rough set is the classification of the domain of interest into disjoint categories [4]. Such a classification refers to the ability to characterize all the classes in a domain. It leads to creation decision table, which can be written in the form IF .. THEN .. rules. Rules raised in this manner can be converted into linguistic form, that are easy readable and interpretable by the unskilled man.

The goal of the task solved in the paper is finding the group of these taxpayers, who evade obligation of paying taxes or trying to decrease value of tax. For this aim the rough set application is used, applying the algorithm classification, which executes process of creation of rule sets describing different data classes. On the basis of input data describing features that represent "clients" and known decision attribute that informs about problems with tax law – all these called information system – there will be created a model, which will be later applied to the new group of input data. This will allow to classify the data with more accuracy, respecting the taxpayer have or have not problems with taxation. Additionally recording rules in linguistic manner will allow tax clerks easier interpret and understand the results.

2. Rough set for classification

In rough sets for classification tasks we define information system [5], where attribute Q is divided into two separable sets, C as conditional attribute set, and D as decision attribute, such $C \cup D = Q$ and $C \cap D$ = <EMPTY SET>. For our task C represents features of tax payers, and D is decision about positive or negative control. Decision table has the following format: DT = $<U, C \cup D, V, f>$, where: U – universe; C, D – as described above; $V = U_{q\in C\cup D}V_q$, where V_q is the set of discrete values of the attribute $q \in Q$; f: U x $(C \cup D) \rightarrow V$ is the decision function, such as $f(x,q) \in V_q$ for every $q \in Q$ and $x \in V$. Decision tables can be divided into deterministic ones where decision attribute value is uniquely specified by combination of conditional attributes, and non-deterministic ones which have opposite meaning. Decision table for our task can be represented in the following form: AGE=MIDDLE, INCOME=HIGH, TAX_RELIEF=HIGH → CON_RESULT=1.

For measuring confidence of decision attributes [1] we need to define dependency coefficient $0 \le k \le 1$, where set of attributes D depends on set on attributes C in a degree defined by k:

$$k = \gamma(C,D) = \frac{card(POS_c(D))}{card(U)}, \qquad (1)$$

where $POS_C(D)$ is a positive region of the partition U/D with respect to C, *card* is the cardinality that is the number of elements in a set.

Then we compute significance of condition $\sigma(a)$ of attribute a, defined by following equation:

$$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D) - \gamma(C-\{a\},D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-\{a\},D)}{\gamma(C,D)}, \qquad (2)$$

Value for significance is $0 \le \sigma(a) \le 1$. If this coefficient equals zero or is beneath fixed threshold for a certain task, this means that the attribute *a* should be eliminated from the computation of decision table, because it has no influence on decision attribute.

3. Research model description

The research process of classification with the use of rough sets theory consists of the six steps, that are represented in the Fig. 1.

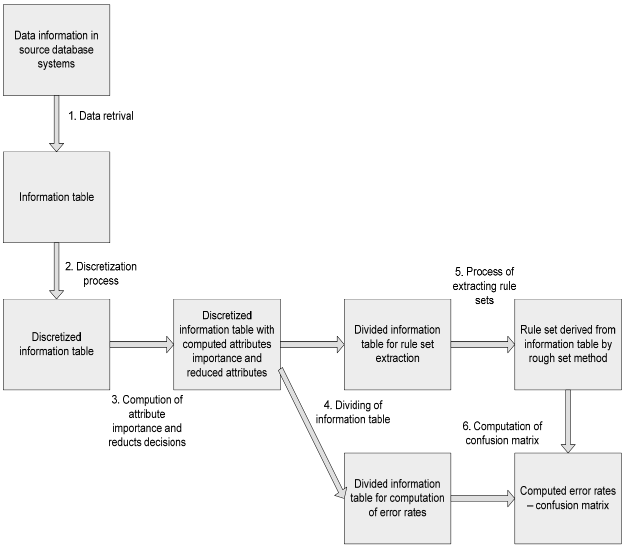
Data information in source database systems
1. Data retrival
Information table
2. Discretization process
Discretized information table
3. Compution of attribute importance and reducts decisions
Discretized information table with computed attributes importance and reduced attributes
4. Dividing of information table
Divided information table for rule set extraction
5. Process of extracting rule sets
Rule set derived from information table by rough set method
6. Computation of confusion matrix
Divided information table for computation of error rates
Computed error rates – confusion matrix

Fig. 1. Research model for classification
Rys. 1. Model badawczy dla zadania klasyfikacji

Data retrival

Gathering the most of possible data about every case for decision system for taxpayers, is very important factor, because of quality of end results of data mining process. On this stage there should not be eliminated any of the input attributes, because this task will be performed in the next steps. Data in form of attributes representing features of taxpayers were retrieved from tax systems by the usage of imp,exp tools, and SQL, and PL/SQL Language, described in [10]. Information table for the research has got the following format (SQL Language):

```
CREATE TABLE PODATNIK
( ID NUMBER,
TOWN VARCHAR2(40),
MARTIAL_S CHAR(1),
GENDER CHAR(1),
AGE NUMBER,
INCOME NUMBER,
TAX_RELIEF NUMBER,
CON_RESULT NUMBER)
```

where,
id – unique identification of taxpayer,
town – place of residence,
matrial_s – martial status of taxpayer,
gender – gender of taxpayer,
age – age of taxpayer,
income – income of taxpayer,
tax_relief – greatness of tax relief,
con_result – result of tax control as decision attribute.

Discretization process

Data discretization is important preparation step for data mining algorithms, because of receiving reasonable number of decision rules, speed of process and inability or deformation of end results because of missing of continuous values of the data [7, 8]. The discretized decision table has eight numeric values of the following meaning:

id – the same as above,
q1 – small, middle or big town,
q2 – not married – married,
q3 – woman – man,
q4 – young, mid, older aged,
q5 – low, mid, high income,
q6 – low, mid, high tax relief,
d1 – negative, positive tax control result.

Computation of attribute importance and decisions concerning the reduction

Calculations was carried out on the basis of rough set theory on the whole discretized decision table as computation of *k*-dependency coefficient, and σ significance of attribute. The decision of reduction of attributes was made on this basis.

Dividing of information table

Information table was divided into two parts in the proportion 60% to 40%: the first was designated for extraction of rule set as decision table, the second for testing purposes of rules. Technologically two views of one table were created in the database.

Process of extracting rule sets

On the basis of rough set theory the rules where extracted [2] and saved in the form of decision table, which values of the input attributes indicate the positive or negative results of tax control. Additionally the information about quantity and strength of certain rule were computed [6].

Computation of Confusion Matrix

At the end of whole process follows the verification of accuracy of the method. For this aim the confusion matrix [11] is computed on the basis of rules and information table for testing purposes. The matrix represents the reliability of results, providing true positive-negative and false positive-negative indicating accuracy of positive predictive values, and negative predictive values of results as well as overall accuracy.

4. Results of experiment

The data from source databases were loaded into information table in Oracle database, and the process of discretization was made. As a result the Table 1 was obtained.

Tab. 1. Form of discretized information table for rule sets extraction
Tab. 1. Tablica informacyjna z danymi dyskretnymi przygotowana do ekstrakcji zbiorów reguł

ID	Q1	Q2	Q3	Q4	Q5	Q6	D1
100001	3	1	1	3	2	2	0
100002	3	1	2	2	2	3	0
100003	3	1	2	2	2	2	0
100004	3	2	2	2	3	3	1
…	…	…	…	…	…	…	…

Then, for every input attribute Q1,Q2,Q3,Q4,Q5,Q6 *k*-dependency coefficient and σ significance of attribute and also global dependency for all attributes were computed. The results are shown in the Table 2.

Tab. 2. Results of dependency and significance of attributes for discretized information table
Tab. 2. Wyniki przedstawiające współczynniki jakości przybliżenia oraz istotności atrybutów dla tabeli informacyjnej z danymi dyskretnymi

Results	Attributes						Global dependency (for all attributes) k
	Q1	Q2	Q3	Q4	Q5	Q6	
k (dependency)	0,0811	0,1203	0,0674	0,0540	0,0970	0,0485	
σ (significance of attribute)	0,3482	0,0327	0,4583	0,5654	0,2202	0,6101	0,1244

Because every input attribute has the value of $\sigma \neq 0$ the reduction of none of attributes where carried out, and all of them will be considered for computation of decision table. From the result we can conduct, that the most import influence on decision attribute have Q6 – TAX_RELIEF.

The result of computation of full decision table is given in Tab. 3.

Tab. 3. Full decision table for described rule sets
Tab. 3. Pełna tablica decyzyjna dla opisanego zbioru reguł

ID	Q1	Q2	Q3	Q4	Q5	Q6	D1	Quantity	Strength
1	3	1	1	1	2	1	0	104	1
2	3	1	2	2	1	1	0	21	1
3	3	2	2	2	3	3	1	198	0.7279
4	3	2	2	3	3	3	1	35	0,6862
…	…	…	…	…	…	…	…	…	…

The simplified decision table as the end result of searching for interested taxpayers is shown in the Table 4.

Tab. 4. Simplified decision table for described rule sets
Tab. 4. Tablica decyzyjna z uproszczonymi regułami decyzyjnymi

ID	Q1	Q2	Q3	Q4	Q5	Q6	D1	Quantity	Strength
1	3	1	2	-	2	-	0	579	0,9136
2	-	2	-	2	-	3	1	347	0,6887
3	-	-	-	2	3	3	1	342	0,6959
4	3	1	-	-	2	3	0	257	0.8754
…	…	…	…	…	…	…	…	…	…

From the Tab. 4 results the method for creating rule sets in the form of IF .. THEN statement. These are easy readable for non-skilled mans. Below the rules are listed:

Rule
No. Description

1. IF TOWN=3 AND MARTIAL_S=1 AND GENDER=2 AND INCOME=2 THEN CON_RESULT=0, Prob = 91,36%
2. IF MARTIAL_S=2 AND AGE=2 AND TAX_RELIEF=3 THEN CON_RESULT=1, Prob. = 68,87%
3. IF AGE=2 AND INCOME=3 AND TAX_RELIEF=3 THEN CON_RESULT=1, Prob. = 69,59%
4. IF TOWN=3 AND MARTIAL_S=1 AND INCOME=2 AND TAX_RELIEF=3 THEN CON_RESULT=0, Prob.=87,54%

From above rules the most interesting one is the rule no 3, where the probability of positive result of tax control equals 69,59%. The rule can be expressed linguistically, in the following form: If taxpayer is middle aged and his/her income is high and his/her tax relief is high, then probability of positive tax control equals 69,59%. From the above described rules there is a fast retrieval of taxpayers that should be controlled.

At the end there is representation of error rates in the form of confusion matrix (Tab. 5).

Tab. 5. Confusion matrix for error rates measures
Tab. 5. Matryca niepewności z wynikami błędów metody

		Predicted data				Overall accuracy
Actual data		1	0	Quantity	Accuracy	
	1	382	56	438	0.872	
	0	457	905	1362	0.664	
	True positive rate	0.46	0.94			0.715

From the matrix it results that the results of the research are reasonable, because accuracy of the most important for end effect positive control results equals 87,2%, and overall accuracy is 71,5%, which are reasonable good results.

5. Software solution

For the task solution the software package in the form of package body of Oracle PL/SQL and SQL languages with the use of intermediate objects on the basis of oracle collection and tables [9] was developed and used. The package computes approximations, dependency coefficients, significance of attributes as well as deterministic and non-deterministic decision table. It works for Oracle 10g and above, also on freely available Oracle 10g Express Edition.

6. Conclusion

The computation of rule sets based on rough sets theory, implemented for Oracle databases can be used with good accuracy results for the task of tax payers classification. It allows to point out the group of tax payer, caused potential problems with obeying of tax law. From the rules described in the form of decision table results a fast and effective method for creating rules in the form IF .. THEN. Next the rules can be written in the linguistic form. This form is easy to interpretation for tax clerks. Through the presented solution the taxpayers choice for the taxation control becomes more accurate, what increases effectiveness of enforcing tax law.

7. References

[1] Pawlak Z.: Some issues on rough sets, Springer Science, 2005.
[2] Inuiguchi M.: Generalizations of Rough Sets and Rule Extraction, Springer, 2005.
[3] Triantaphyllou E., Felici G.: Data Mining & Knowledge Discovery based in Rule Induction, Springer Science, 2006.
[4] Pawlak Z.: Rough Sets – Theoretical Aspects of Reasoning about Data., Kluwer Academic Publishers, 1991.
[5] Olson D., Delen D.: Advanced Data Mining Techniques, Springer, 2008.
[6] Duntsch I., Gediga G.: Rough set data analysis, Methodos Publisher, 2000.
[7] Bazan J., Synak P., Wrobleski J.: Rough Set Algorithms in Classification Problem, Springer, 2000.
[8] Grzymala-Busse J.: Rough Set Strategies to Data with Missing Attribute Values, Springer, 2006.
[9] Oracle Database 10g: Advanced PL/SQL, Oracle corp., 2004.
[10] Oracle Database Documentation Library 11g Release 1 (11.1), Oracle Corp., 2008.
[11] Oracle Data Mining Concepts 11g Release 1 (11.1), Oracle Corp., 2005-2007.
[12] Elmasri R., Shamkant B.Navathe, Fundamentals of Database Systems, Addison-Wesley, 2004.

Artykuł recenzowany