

Business Decision Making by Systematic Syntactic Classification of Objects

Visnja Ognjenovic*, Eleonora Brtko*, Martin Jovanovic**, Vladimir Brtko*, Ivana Berkovic*

* University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

** University of Niš, Faculty of Electronic Engineering, Niš, Serbia

visnjao@tfzr.uns.ac.rs, brtkav@gmail.com, martin.jovanovic@elfak.ni.ac.rs

Abstract – This paper presents a possible way of business decision making by Systematic Syntactic Classification of Objects (SSCO). Possibilities that SSCO system offers, as well as its relation with rough sets, are described. Well-known German Credit dataset has been used as an example for business decision analysis. Based on the given data on a credit applicant, each applicant has been marked as "good" or "bad". SSCO system yielded relations through IF-THEN rules which enable the classification of all applicants. A comparison to other systems, also using German Credit dataset, has been made.

I. INTRODUCTION

In the process of business decision making many factors are interlaced such as attribute selection, parameter selection, selection of the type of analysis, dataset selection etc. The business analysis can be conducted by many different approaches. Because of the complexity of contemporary business decision making, the algorithms from the domain of artificial intelligence are used. There are many different approaches such as: neural nets, decision trees, rough sets, graph-based relational concepts and many others. Every particular problem which has to be solved needs a convenient approach.

For business decision making process in small and medium sized enterprises (SMEs) when domain experts are not available, software tools based on the artificial intelligence algorithms are very useful and cheap solution. The main tasks where artificial intelligence tools are commonly used are [1]: Approximation, Optimization, Classification, Prediction, Generalization, Relation-making, Abstraction, Adaptiveness. In this paper will be shown how to use artificial intelligence tools as well as soft computing tools based on SSCO [2,3] for: Prediction, Generalization and Relation-making. The prediction of the output data is achieved according to the inputs which are continuous or discrete. Outputs are continuous only In this property lies the main difference between prediction and classification. Generalization is used when according to a training set some general patterns are extracted and used to classify unseen object. As was mentioned in [1], general patterns enable more precise classification or prediction. Relation-making process defines relations between input and output as well as between different attributes. This approach is used to make a groups of input data as well as a relations in each group.

The paper deals with the possibilities of business decision making by usage of the SSCO system. The German Credit data [4] is used as an example data set in order to decide which attributes are most relevant to decision making process. Based on the given data on a credit applicant, each applicant has been marked as "good" or "bad". The SSCO system generated decision rules in the IF THEN form.

In second section of this paper, theoretical bases of the SSCO system are described The linkage between SSCO and the rough sets theory is underlined The third section contains description of the German Credit data. Fourth section contains the results of the SSCO when applied to German Credit dataset. In fifth section a comparison between SSCO achieved results and results achieved by some other systems is given. The conclusion is given in a section six.

II. SSCO SYSTEM

The SSCO system was developed in the 2005 – 2007 period as a result of doctoral research of V. Brtko [2,5]. The theoretical fundaments are based on the rough sets theory as a part of broader domain of the soft computing and on classical theory of the state space search. Both theories are very useful for automated decision rules synthesis. SSCO algorithm is an original approach which is compatible with the rough sets theory in many applications [6].

A. Rough sets theory

The rough sets theory was originally discovered by Zdzisław I. Pawlak in 1980s. He has developed the mathematical tool to cope with ambiguous and incomplete data, as well as with the vague concepts. As the mathematical basis of this theory, there is an indiscernibility relation defined on an information system in the context of the rough sets [7]. Let U be a universe (finite set of objects), $Q = \{q_1, q_2, \dots, q_m\}$ is a finite set of attributes, V_q is the domain of attribute q and $V = \bigcup_{q \in Q} V_q$.

Definition 1. *Information System.* An information system is the quadruple $S = \langle U, Q, V, f \rangle$ where $f = U \times Q \rightarrow V$ is a total function such that $f(x, q) \in V_q$ for each $q \in Q, x \in U$, called information function.

If some of the attributes are interpreted as outcomes of classification, then the information system $S = \langle U, Q, V, f \rangle$ can be defined as a decision system by $DS = \langle U, C, D, V, f \rangle$, where $C \cup D = Q$, $C \cap D = \emptyset$. Set C is called the set of condition attributes and set D is called the set of decision attributes [7]. In practice, there is one decision attribute.

Definition 2. Indiscernibility Relation. To every non-empty subset of attributes P is associated an indiscernibility relation on U , denoted by I_P :

$$I_P = \{(x, y) \in U \times U : f(x, q) = f(y, q), \forall q \in P\} \quad (1)$$

The relation (1) is an equivalence relation – reflexive, symmetric and transitive. The family of all the equivalence classes of the I_P is denoted by U/I_P and class containing an element x by $I_P(x)$ [8].

Definition 3. Set approximations. Let X be a non-empty set of U and $\emptyset \neq P \subseteq Q$.

Set X is approximated by means of P-lower (2) and P-upper (3) [8] approximations of X :

$$\underline{P}(X) = \{x \in U : I_P(x) \subseteq X\} \quad (2)$$

$$\overline{P}(X) = \bigcup_{x \in X} I_P(x) \quad (3)$$

The P-boundary of X is denoted by $Bn(X)$:

$$Bn(X) = \overline{P}(X) - \underline{P}(X) \quad (4)$$

One natural dimension of reducing data is to identify equivalence classes. This could be achieved by keeping only those attributes that preserve the indiscernibility relation and consequently, set approximation. The rejected attributes are redundant (superfluous) since their removal cannot worsen the classification. Let $\emptyset \neq P \subseteq Q$ and $a \in P$. Attribute a is superfluous in P if $I_P = I_{P-\{a\}}$, otherwise it is indispensable attribute. The set P is orthogonal if all its attributes are indispensable. The set $P - \{a\}$ is a reduct of P if it is orthogonal and $I_P = I_{P-\{a\}}$ [6].

It is possible to find the minimal set of consistent decision rules (logical implications) that characterize some sample system which describes an information system. It is possible to investigate rules of the form: IF α THEN β . Here α (rule's antecedent) denotes a conjunction (AND logical operator) of descriptors that only involve attributes of some reduct and let β (rule's consequent) denote a descriptor $d = v$, where d is decision attribute and v is allowed decision value [6].

Reduct set is not unique in every situation, there is a possibility that multiple reduct sets are calculated. It is very important to find minimal (shortest) reduct set. Minimal reduct set produces shortest decision rules.

The length of generated rules is directly proportional to used reduct set: all attributes from reduct are used in the IF part of each rule [2].

B. The description of the SSCO system

There are multiple variations of the SSCO system. In this research the version 3.4 is used. There are standalone versions of the SSCO system as well as the version which was incorporated to a web based system in order to enable functionality for data analysis (www.tfzr.uns.ac.rs/dawp).

The main menu have four items: File, Task, Output and About. File menu item is shown on Figure 1. This menu item enable the loading of the input data and output of the generated rules in the matrix form.

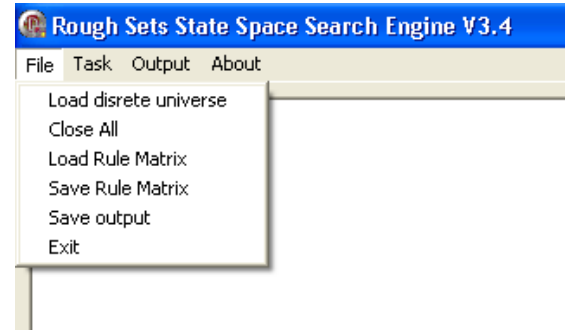


Figure 1. SSCO system – File option

A special feature that SSCO system offers is to generate a small number of rules, which in practice allows easier and faster manipulation. This is well demonstrated in working with medical data [2, 5]. The Task option is shown on Figure 2. This option enables a choice of the format in which results of the analysis are presented.

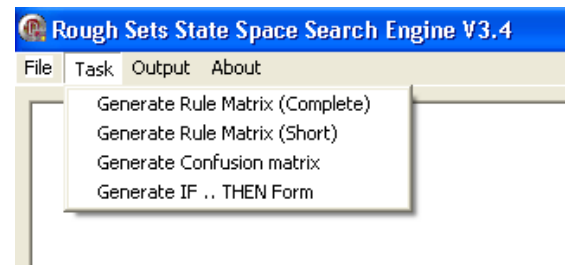


Figure 2. SSCO system – Task option

The results of the SSCO system are in the IF THEN form accompanied by the probability measure, so that results can be evaluated accordingly. In the IF THEN rule:

[4, 0.75] IF (a1, 1), (a2, 6), (a3, 4), (a4, 1) THEN (a6, 1)

a pair [4, 0.75] represents a number of the objects which support the rule and probability of the rule.

The IF part of the rule contains the condition attributes and their corresponding values while THEN part of the

rule contains decision attribute and the corresponding value.

The previously described rule set is accompanied by so called confusion matrix. Confusion matrix C is a $|V_d| \times |V_d|$ matrix, where V_d is the set of possible values of decision attribute. This matrix with integer entries summarizes the performance of rule set while classifying the set of objects. Entry:

$$C_{i,j} = |\{x \in U : d(x) = i, \bar{d}(x) = j\}|,$$

where $d(x)$ is the actual decision and $\bar{d}(x)$ is the predicted decision, which counts the number of objects that really belong to class i , but were classified to class j . The classification percentage is calculated based on a ratio of the number of diagonal elements of the matrix (sum of the diagonal elements) and total number of the objects.

It is important to notice that confusion matrix generated by SSCO system was formed without any voting system so that one object could be classified to more than one class by some inexact rules.

III. GERMAN CREDIT DATA

German Credit data is a well known dataset which was donated by Professor Dr. Hans Hofmann from Institut für Statistik und Ökonometrie, University of Hamburg [4]. Number of Instances is 1000. Two datasets are provided: the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes. For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric", which contains 24 condition attributes and one decision attribute. Table I shows the attributes and their values.

TABLE I. ATTRIBUTES OF GERMAN CREDIT DATASET

Attribute	Type	Values
Checking account	qualitative	1: < 0 DM; 2: \$ 0 and < 200 DM; 3: \$ 200 DM/salary assignments for at least one year; 4: no checking account
Duration (in months)	numerical	
Credit history	qualitative	0: no credits taken/all credits paid back duly; 1: all credits at this bank paid back duly; 2: existing credits paid back duly till now; 3: delay in paying off in the past; 4: critical account/ other credits existing (not at this bank)
Purpose	qualitative	0: car (new); 1: car (old); 2: furniture/equipment; 3: radio/television; 4: domestic appliances; 5: repairs; 6: education; 7: vacation; 8: retraining; 9: business; 10: others
Credit amount	numerical	
Savings account	qualitative	1: < 100 DM; 2: \$100 and < 500 DM; 3: \$500 and < 1000 DM; 4: \$1000 DM; 5: unknown/no savings account
Present	qualitative	1: unemployed; 2: <1 year; 3:

employment since		\$1 and < 4 years; 4: \$4 and < 7 years; 5: \$7 years
Installment rate in percentage of disposable income	numerical	
Personal status and sex	qualitative	1: male, divorced/separated; 2: female, divorced/separated/married; 3: male, single; 4: male, married/widowed; 5: female, single
Other parties	qualitative	1: none; 2: co-applicant; 3: guarantor
Present residence since	numerical	
Property	qualitative	1: real estate; 2: if not 1: building society savings agreement/ life insurance; 3: if not 1/2 : car or other; 4: unknown/ no property
Age in years	numerical	
Other installment plans	qualitative	1: bank; 2: stores; 3: none
Housing	qualitative	1: rent; 2: own; 3: for free
Number of existing credits at this bank	numerical	
Job	qualitative	1: unemployed/unskilled-non-resident; 2: unskilled-resident; 3: skilled employee/official; 4: management/self-employed/ highly qualified employee/officer
Number of people being liable to provide maintenance for	numerical	
Telephone	qualitative	1: none; 2: yes, registered under the customers name
Foreign worker	qualitative	1: yes; 2: no
Score of the applicant	qualitative	1: Good; 2: Bad

This dataset requires use of a cost matrix:

	1	2

1	0	1

2	5	0

where 1 = Good, 2 = Bad. The rows represent the actual classification and the columns represent the predicted classification.

So, it is much worse to classify a customer as good when he is bad (5), than to classify a customer as bad when he is good (1).

IV. RESULTS

In the data analysis process all condition attributes were analyzed in order to extract those which have a greatest impact to a decision. Following attributes have been extracted:

- Status of existing checking account

- Duration
- Credit history
- Savings account
- Other installment plans

The objects of the universe are separated to two sets: training set (500 objects) and test set (500 objects).

According to training set, SSCO system produced 350 rules in the IF THEN form:

```
IF THEN Form
[1,1]      IF (a1,4), (a2,60),
(a3,4) THEN (a6,-1)
[1,1]      IF (a1,4), (a2,60),
(a3,2) THEN (a6,-1)
[1,1]      IF (a1,4), (a2,54) THEN
(a6,-1)
```

...

```
[1,1]      IF (a1,1), (a2,6),
(a3,4), (a4,3) THEN (a6,-1)
[4,0.25]   IF (a1,1), (a2,6),
(a3,4), (a4,1) THEN (a6,2)
[4,0.75]   IF (a1,1), (a2,6),
(a3,4), (a4,1) THEN (a6,1)
[2,1]      IF (a1,1), (a2,6),
(a3,2), (a4,3) THEN (a6,-1)
[2,0.5]    IF (a1,1), (a2,6),
(a3,2), (a4,1), (a5,3) THEN (a6,2)
[2,0.5]    IF (a1,1), (a2,6),
(a3,2), (a4,1), (a5,3) THEN (a6,1)
[1,1]      IF (a1,1), (a2,6),
(a3,2), (a4,1), (a5,1) THEN (a6,-1)
```

The objects from the test set are classified by these rules, so that following confusion matrix has been created:

Confusion matrix:

	1	2
1	236	90
2	96	83

249 objects are classified by exact rules, 128 objects classified by inexact rules, 123 objects are not classified

The SSCO system has correctly classified 75,4 % of objects.

Further test was conducted on two attributes:

- Status of existing checking account
- 1-good, 2-bad

in order to compare the SSCO results with the results of some other systems. According to training set, SSCO system has generated following rules in the IF THEN form:

```
IF THEN Form
[197,0.091] IF (a1,4) THEN (a2,2)
[197,0.908] IF (a1,4) THEN (a2,1)
[31,0.193]  IF (a1,3) THEN (a2,2)
[31,0.806]  IF (a1,3) THEN (a2,1)
[144,0.402] IF (a1,2) THEN (a2,2)
[144,0.597] IF (a1,2) THEN (a2,1)
[128,0.421] IF (a1,1) THEN (a2,2)
[128,0.578] IF (a1,1) THEN (a2,1)
```

For further classification process of the test set, two most reliable rules were separated:

```
[197,0.908] IF (a1,4) THEN (a2,1)
[31,0.806]  IF (a1,3) THEN (a2,1)
```

These rules mean:

```
if Checking account = 4 then
Applicant = good
```

```
if Checking account = 3 then
Applicant = good
```

The confusion matrix was generated from test set by two these rules:

Confusion matrix:

	1	2
1	336	336
2	164	164

0 objects are classified by exact rules, 500 objects classified by inexact rules, 0 objects are not classified

This confusion matrix approves reliability of the selected rules.

V. COMPARISON

The comparison is done by the insight to the results published in [9] for the same German Credit data. As in [9] decision rules were generated by multiple algorithms, so training set produced following rules:

```
if (Checking account ≥ 3) then
Applicant = good
```

```
if (Duration = 1) and (Other
installment plans ≥ 2) then Applicant =
good
```

```

if (Other installment plans  $\geq$  2) and
(Credit history = 4) then Applicant =
good

```

```

if (Duration = 1) and (Credit
history = 4) then Applicant = good

```

```

if (Savings account  $\geq$  3) then
Applicant = good

```

```

if (Other parties = 3) then
Applicant = good

```

Table II presents the results of C4.5, C4.5rules, Neurorule, Trepan, and Nefclass to the discretized data sets. Results that are obtained are:

TABLE II. RESULTS (DIFFERENT ALGORITHMS)

Data Set	Algorithm	Train (%)	Test (%)
German credit	C4.5	80.63	71.56
	C4.5 rules	81.38	74.25
	Neurorule	76.13	75.15
	Trepan	75.38	73.95
	Nefclass	73.57	73.65

Attributes selected by several different algorithms were also selected by SSCO system which is very important. Some rules generated by SSCO system can be merged to one rule:

```

if Checking account  $\geq$  3 then
Applicant = good

```

This rule can also be find in the resulting decision rules list generated by several different algorithms. Other relevant attributes can be used to check the relevance of other decision rules. This is a confirmation of the relevance of the attributes: Status of existing checking account, Duration, Credit history, Savings account and Other installment plans which were selected by SSCO system.

In [10] GP algorithm was compared to C4.5 algorithm, similar results were obtained. Some results are shown in Table III:

TABLE III. GP AND C4.5 COMPARISON

Data Set	Algorithm	Average
German credit	simple GP	27.1
	C4.5	27.2

According to this, we can conclude that SSCO system is applicable for tasks of classification and prediction of the German Credit data.

VI. CONCLUSION

This paper deals with the description of the SSCO data analysis system which is based on theoretical domains of

artificial intelligence and soft computing (rough sets theory and state space search) It is shown how to use SSCO system for classification and prediction tasks. Both, classification and prediction are very important in the business decision making process. The SSCO system was described. The performance of the system was tested on German Credit data. It is shown how to operate with multiple attributes in business decision making process.

The results obtained by SSCO system are compared to the results obtained by some other systems on the same German Credit data. It is shown how SSCO system can be used for decision rules synthesis. These rules are accompanied by probability measure. The results of the SSCO system largely coincide with decision rules generated by other systems. This shows a quality of the SSCO system.

Finally, we can conclude that systems for the analysis of data, such as SSCO system, are very useful in the domain of business decision making. These systems are very useful for small and medium sized enterprises because they are cheap and easy to implement and use.

REFERENCES

- [1] E. Y. Li, "Artificial neural networks and their business applications", Information & Management, Volume 27, Issue 5, November 1994, Pages 303-313.
- [2] V. Brtko, E. Stokic, B. Srdic, "Automated extraction of decision rules for leptin dynamics—A rough sets approach", Journal of Biomedical Informatics 41, pp. 667 – 674, 2008.
- [3] V. Brtko, I. Berkovic, E. Stokic, B. Srdic, "A comparison of rule sets generated from Databases by indiscernibility relation - A rough sets approach", ICCP 2007: IEEE 3rd International Conference on Intelligent Computer Communication and Processing, Proceedings, (2007), vol. br. , str. 279-282.
- [4] C. L. Blake, C. J. Merz: UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/>
- [5] V. Brtko, I. Berkovic, E. Stokic, B. Srdic, "Automated extraction of decision rules from medical databases - A rough sets approach" (Proceedings Paper), 2007 5TH INTERNATIONAL SYMPOSIUM ON INTELLIGENT SYSTEMS & INFORMATICS, (2007), vol. br. , str. 27-3
- [6] V. Brtko, I. Berkovic, E. Brtko, V. Jevtic, "A Comparison of Rule Sets Induced by Techniques Based on Rough Set Theory", SISY 2008, 6th International Symposium on Intelligent Systems and Informatics, September 26-27, 2008 Subotica, Serbia, IEEE Catalog Number: CFP0884C-CDR, ISBN: 978-1-4244-2407-8, Library of Congress: 2008903275.
- [7] Z. Pawlak, "Rough sets: Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, Dordrecht (1991)
- [8] J. Komorowski, Z. Pawlak, L. Polkowski, A. Skowron, "Rough Sets: A Tutorial", <http://citeseer.ist.psu.edu/komorowski98rough.html>, 1998
- [9] B. Baesens, R. Setiono, C. Mues, S. Viaene, J. Vanthienen, "Building Credit-Risk Evaluation Expert Systems. Using Neural Network Rule Extraction and Decision Table", Proceedings of the International Conference on Information Systems, ICIS 2001, AIS Electronic Library (AISeL)
- [10] J. Eggermont, J.N. Kok, and W.A. Kusters, "Genetic Programming for data classification: partitioning the search space", , SAC'04, March 14–17, 2004, Nicosia, Cyprus