Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 179 – 187
ISSN 1896-7094

Credibility Coefficients Based on Decision Rules

Roman Podraza[1], Mariusz Walkiewicz[1], Andrzej Dominik[2]

[1]Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
`{R.Podraza, M.Walkiewicz}@elka.pw.edu.pl`
[2]Institute of Radioelectronics, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
`A.Dominik@elka.pw.edu.pl`

Abstract. Credibility coefficients are heuristic measures evaluating similarity of objects in respect to other data in information systems (or decision tables). By applying knowledge discovery methods it is possible to reveal rules from the information system. However the knowledge obtained from the data can be not precise due to improper data. It is assumed that majority of data is correct and only a minor part may be improper. Credibility coefficients of objects are aimed to indicate to which group a particular object probably belongs. A main focus of the paper is set on an algorithm of calculating credibility coefficients. This algorithm is based on decision rules, which are generated using the rough set theory. Implementation and applications of credibility coefficients are presented in the paper. Discussion of some practical results of identifying improper data by credibility coefficients is inserted as well.

1 Introduction

Credibility coefficients [1, 2, 3, 4] were introduced to identify improper objects in information systems or decision tables. Credibility coefficients are defined as a heuristic measure from range <0.0; 1.0>, where values close to the lower bound denote a low credibility, whereas numbers close to upper bound denote a high credibility. The whole concept is based on a supposition that majority of data is trustworthy and only minority of it can be considered as corrupted, improper or exceptional. Calculations of credibility coefficients evaluate similarities between data.

The ARES Rough Set Exploration System [1, 5] is a data analysis tool based on the rough set theory [6, 7, 8]. Its functionality includes a vast data analysis tools leading to discovering rules by applying different algorithms. A unique feature of the ARES System is a possibility to evaluate credibility coefficients for objects from decision tables. Some algorithms were already published [1, 3, 4] and this paper presents an approach based on decision rules deduced from the decision table.

The ARES Rough Set Exploration System is a general data analysis tool, but it was designed and developed for medical applications [9, 10]. In medicine and other natural sciences exceptions to the rules may be more interesting than the rules themselves. For instance, it is very important to identify a disease when symptoms are misleading, when a case does not fit to the rules. Skills of a good physician include a way of distinguishing and dealing with exceptional cases. The main purpose of introducing credibility coefficients was to provide an automatic aid in expert systems for identifying exceptional cases to draw a special attention of specialists to these cases.

The paper comprises a very short description of rough set theory to enable expressing mathematical descriptions of credibility coefficients based on decision rules. An intuitive description and explanation of the algorithm is given as well. The next chapters present an example of a tiny decision table, for which the credibility coefficients were computed, and a proposal of developing the algorithm. The paper is completed with some conclusions and suggestions how credibility coefficients can be applied in practice.

2 Rough Set Concepts

Rough set theory can be applied for analyzing data in an information system. The information system S can be defined as $S = <U, Q, V, f>$, where U is a finite set of objects, Q is a finite set of attributes, $V = \sum_{q\in Q} V_q$ and V_q is a domain of the attribute q and $f\colon U\times Q \rightarrow V$ is a function that $f(x, q)\in V_q$ for every $x\in U$, $q\in Q$.

An information system can be represented by a table, where rows correspond to objects and columns correspond to attributes. Every cell stores a value of the given attribute for a particular object.

An information system can be regarded as decision table if the set of all attributes is split into condition attributes C and decision attributes D ($Q = C\cup D$ and $C\cap D=\emptyset$). Information system $S = <U, C\cup D, V, f>$ is deterministic iff $C\rightarrow D$; otherwise is non-deterministic.

For further consideration we assume that number of decision attributes is limited to one. This restriction is often met in practical data analysis tools and the ARES Rough Set Exploration System, which was used to perform some analysis, has such feature as well.

Elementary condition is a pair of attribute-value. Every object is represented or satisfies a set of elementary conditions represented by cells of information system (or decision table). Set of all elementary conditions of object $t\in U$ is denoted as *Inf(t).*

Coverage of set of elementary conditions P (denoted as ‹P›) in a given information system is a set of objects satisfying all conditions represented by P.

Support of set of elementary conditions P (denoted as *sup(P)*) in a given information system is a cardinality of set ‹P›, which is a number of objects satisfying all conditions represented by P.

A set of elementary conditions is called a frequent set if its support is greater (or greater-equal) than a given value.

In rule $t\rightarrow s$ antecedent t is a conjunction of elementary conditions from condition attributes and consequent s is elementary condition of the only decision attribute.

A confidence of rule $t\rightarrow s$ is a quotient of supports of the rule and of its antecedent, respectively.

$$conf(t \rightarrow s) = \frac{\sup(t \wedge s)}{\sup(t)}$$

3 Algorithm of Calculating Credibility Coefficients

3.1 Descriptions

The following notation is used in describing the algorithm of calculating credibility coefficients based on decision rules:

- *W[]* – vector *W,* which index domain may be any set of data, in particular for object $t\in DT$, *W[t]* denotes value of vector element, which is associated with object *t* (e.g. vectors *counts[], decCount[], CFS[]*),
- *Inf(t)* – set of elementary conditions based on values of successive attributes of object *t*

- $(X \rightarrow Y).conf$ – confidence of rule $X \rightarrow Y$

3.2 Algorithm to Compute Credibility Coefficients

Input data:

- AR - set of decision rules
- DT - decision table

Output data:

- C_R[]-- vector of credibility coefficients for all objects

```
1   counts[] = New counts[]
2   C_R'[] = New C_R'[]
3   Forall (X → Y) ∈ AR Do
4      Forall t ∈ DT Do
5         If X ⊂ Inf(t) And Y ∉ Inf(t) Then
6            counts[t] := counts[t] + 1
7            C_R'[t] := C_R'[t] + (X → Y).conf
8   Forall t ∈ DT Do
9      C_R[t] := 1
10     If (count[t] <> 0) Then
11        C_R[t] := C_R[t] – C_R'[t] / counts[t]
```

For all rules the algorithm investigates all objects form the decision table (lines 3-7). For objects, which satisfy an antecedent of the analyzed rule and at the same time have decision values different than the decision of the rule, an auxiliary value of credibility coefficient ($C_R[t]$) is incremented by a confidence of the rule (not supported by the object) and a counter for the object is incremented by one.

The last part of the algorithm (lines 8-11) sets initial values of the credibility coefficients to maximum (1). If a counter of a particular object is not zero (it means, there was at least one rule, which antecedent was satisfied by the object and decision values of the objects and the rule were different) the value of the credibility coefficient is decremented by the quotient of the auxiliary value and the counter.

The idea of the algorithm is to punish such objects, which do not fit to the rules of the system. The objects exposed by the algorithm (by assigning lower values of credibility coefficients) satisfy only antecedents of rules (one or more), but have their decisions set not in accordance to the rules. The penalty to the maximal value of the credibility coefficient is an average confidence of the violated rules.

More formally the credibility coefficient C_R for object $u \in U$ from a decision table $TD=(U, C \cup \{d\}, V, f)$ and set of rules AR can be presented as

$$C_R(u) = \begin{cases} 1 & for \quad S(u) = \phi \\ 1 - \dfrac{C_R^{'}(u)}{|S(u)|} & for \quad S(u) \neq \phi \end{cases}$$

$$S(u) = \{ \quad (X \to Y) \in AR : u \in \langle X \rangle \ \wedge \ u \notin \langle \{Y\} \rangle \ \}$$

$$C_R^{'}(u) = \sum_{(X \to Y) \in S(u)} (X \to Y).conf$$

The presented algorithm can be slightly modified. It can process only possible rules (and not all rules as presented in the given version). The modification can be implemented in two ways. The first option is to input a set of possible rules for the original algorithm. The second option is to introduce a modification in line 3 of the algorithm checking if a particular rule is a possible one and only such rules should be considered in further processing.

The modification of the algorithm results in a slight change in idea of its meaning. A specific feature of possible rules (but not certain) is used in the proposal. If an object covers the antecedent of the rule and have a different value of the decision it means that there are some other objects which are indiscernible with this one (from the point of view of the condition attributers) and have the decision equal to the consequent of the rule. Such object is punished by decreasing its credibility coefficient by average confidence of rules following the described circumstances.

4 Example

Application of the credibility coefficient is presented on example of six objects representing a group of patients (Table 1). There are three condition attributes (headache, myalgia and temperature) and one decision attribute (flue). Values of all attributes are presented in form of texts and corresponding integer number (coded data in parentheses). The decision table is extended by two columns with values of credibility coefficients based on decision rules. $\mathbf{C_{AR}}$ denotes credibility coefficient evaluated from all rules and $\mathbf{C_{PR}}$ is based on possible rules. Both coefficients were applied to rules with minimal support equal to 33% and minimal confidence equal to 50%.

Table 1 . Credibility coefficients based on decision rules for a set of patients

No.	Headache	Myalgia	Temperature	Flue	$\mathbf{C_{AR}}$	$\mathbf{C_{PR}}$
1	No (0)	Yes (1)	High (0)	Yes (1)	1.00	1.00
2	Yes (1)	No (0)	High (0)	Yes (1)	1.00	1.00
3	Yes (1)	Yes (1)	Very High (1)	Yes (1)	1.00	1.00
4	No (0)	Yes (1)	Very High (1)	Yes (1)	1.00	1.00
5	Yes (1)	No (0)	High (0)	No (0)	0.34	0.34
6	No (0)	Yes (1)	Normal (2)	No (0)	0.31	1.00

Only objects 5 and 6 have credibility coefficients less than 1. Values of credibility coefficient $\mathbf{C_{AR}}$ are consequences of small number of rules applicable to the objects. Value of credibility coefficient $\mathbf{C_{PR}}$ for object 5 indicates that this is the only object for which there are possible rules (at least one), which have antecedents covered by the object and consequents different than decision of the object.

The similar tests were performed for the following rule parameters:

- Minimal support (in number of objects): {1, 2, 3}.
- Minimal confidence (in %): {25, 50, 75, 100}.

The results are presented in Table 2, which contains as well number of rules applicable in each case. Every combination of two values of the support and the confidence is labeled by a variant for further discussion of the results.

Table 2a. Series of credibility coefficients based on decision rules with confidence values equal to 25% and 50% for set of patients from Table 1

Conf.		25						50					
Supp.		1		2		3		1		2		3	
Var.		v1		v2		v3		v4		v5		v6	
|AR|		34		7		1		29		7		1	
|PR|		26		4		0		26		4		0	
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}
Patients	1	0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	2	0,5	0,5	1,0	1,0	1,0	1,0	0,5	0,5	1,0	1,0	1,0	1,0
	3	0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	4	0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	5	0,5	0,5	0,3	0,3	1,0	1,0	0,5	0,5	0,3	0,3	1,0	1,0
	6	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0

Table 2b. Series of credibility coefficients based on decision rules with confidence values equal to 75% and 100% for set of patients from Table 1

Conf.		75						100	
Supp.		1		2		3		1	2
Var.		v7		v8		v9		v10	v11
|AR|		15		3		1		14	2
|PR|		14		2		0		14	2
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{PR}	C_{PR}
Patients	1	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	2	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	3	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	4	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	5	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	6	0,3	1,0	0,3	1,0	0,3	1,0	1,0	1,0

For minimal support set to 3 (variants v3, v6 and v9) coefficient C_{AR} for object 6 has non-maximal value. There were no possible rules for these cases and hence values of coefficient C_{PR} were evaluated to 1. For minimal confidence equal to 100% all rules were possible and certain, so no objects could be punished by algorithm evaluating the credibility coefficients. For minimal support set to 75% (variants v7, v8, v9) only object 6 has non-maximal value for coefficient C_{AR} and all other coefficient values are highest. For variants v1 and v4 objects 2 and 5 get non-maximal value. The better discrimination of objects in these cases (characterized by small support and small confidence for rules involved) is caused by more significant number of the rules and then more precise evaluation of each object.

The next table (Table 3) presents the same series of credibility coefficients when object 5 was removed form the decision table. This object was indicated as the “worst” one by relatively low values of credibility coefficients based on decision rules. If it is the exception to the rules we are curious how works the same algorithm on data without the exception.

For all variants, values of coefficient C_{PR} are utmost, because the new decision table (without object 5) is deterministic one, so all rules are certain. For variants v2-v9 coefficients C_{AR} have non-maximal values only for object 6, which is the only one having decision different than the other objects. For variant v1 all objects but object 2 have non-maximal value of credibility coefficient C_{AR}. The objects were “punished” by rules generated from object 6 (in variant v1 minimal support is 1).

Table 3a. Series of credibility coefficients based on decision rules with confidence values equal to 25% and 50% for set of patients from Table 1 without object 5

Conf.		25						50					
Supp.		1		2		3		1		2		3	
Var.		v1		v2		v3		v4		v5		v6	
|AR|		27		7		1		24		7		1	
|PR|		21		4		0		21		4		0	
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}
Patients	1	0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	2	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	3	0,8	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	4	0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	6	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0

Table 3b. Series of credibility coefficients based on decision rules with confidence values equal to 75% and 100% for set of patients from Table 1 without object 5

Conf.		75						100	
Supp.		1		2		3		1	2
Var.		v7		v8		v9		v10	v11
|AR|		22		5		1		21	2
|PR|		21		4		0		21	2
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}
Patients	1	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	2	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	3	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	4	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	6	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0

5 Modification of the Algorithm

The proposed algorithm for evaluating credibility coefficients based on decision rules has a very important drawback. Initial value of the credibility coefficient for every object is set to one. Only such objects which cover antecedents of the rules can have their initial value of credibility coefficients modified (in minus). Some objects may do not fit to any rules, because their set is limited by values of minimal support and minimal confidence of the rules. Such objects were not involved in calculations, and pretended to be perfectly appropriate in the decision table, which is obviously not true. The objects will be called uncertain and modification of the algorithm is aimed to point them out.

Let us have the modification of the algorithm presented in chapter 3.

```
counts[] = New counts[]
visited[] = New visited[]
C_R'[] = New C_R'[]
Forall (X → Y) ∈ AR Do
   Forall t ∈ DT Do
      If X ⊂ Inf(t) Then
         visited[t] := TRUE
         If Y ⊄ Inf(t) Then
            counts[t] := counts[t] + 1
            C_R'[t] := C_R'[t] + (X → Y).conf
Forall t ∈ DT Do
   If visited[t] = TRUE Then
      C^M_R [t] := 1
   Else
      C^M_R [t] := -1
      If (count[t] <> 0) Then
         C^M_R[t] := C^M_R[t] - C_R'[t] / counts[t]
```

More formally, the modified credibility coefficient C^M_R for object $u \in U$ from decision table $TD = (U, C \cup \{d\}, V, f)$ and set of rules AR can be expressed as below.

$$C_R^M(u) = \begin{cases} -1 & \text{for } W(u) = \phi \\ 1 & \text{for } W(u) \neq \phi \ \wedge \ S(u) = \phi \\ 1 - \dfrac{C_R'(u)}{|S(u)|} & \text{for } S(u) \neq \phi \end{cases}$$

$$W(u) = \{ \ (X \rightarrow Y) \in AR: \ u \in \langle X \rangle \ \}$$

$$S(u) = \{ \ (X \rightarrow Y) \in AR: \ u \in \langle X \rangle \ \wedge \ u \notin \langle \{Y\} \rangle \ \}$$

$$C_R'(u) = \sum_{(X \rightarrow Y) \in S(u)} (X \rightarrow Y).conf$$

Value -1 is a special one for denoting the uncertain objects. It does not belong to the domain of the credibility coefficient and is used only for objects for which the algorithm cannot be properly applied. Such objects may be interested as a different kind of exceptions (in contradiction do exceptions pointed out by the credibility coefficients).

Table 4 presents the impact of the modification on values of credibility coefficients evaluated with the same assumptions as for Table 2. The uncertain object is denoted by '?' – when the algorithm fails in classifying the credibility of the object.

Table 4a. Series of modified credibility coefficients based on decision rules with confidence values equal to 25% and 50% for set of patients from Table 1.

Conf.		25						50					
Supp.		1		2		3		1		2		3	
Var.		v1		v2		v3		v4		v5		v6	
|AR|		34		7		1		29		7		1	
|PR|		26		4		0		26		4		0	
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}
Patients	1	0,7	1,0	1,0	1,0	1,0	?	1,0	1,0	1,0	1,0	1,0	?
	2	0,5	0,5	1,0	1,0	?	?	0,5	0,5	1,0	1,0	?	?
	3	0,7	1,0	1,0	1,0	1,0	?	1,0	1,0	1,0	1,0	1,0	?
	4	0,7	1,0	1,0	1,0	1,0	?	1,0	1,0	1,0	1,0	1,0	?
	5	0,5	0,5	0,3	0,3	?	?	0,5	0,5	0,3	0,3	?	?
	6	0,3	1,0	0,3	?	0,3	?	0,3	1,0	0,3	?	0,3	?

Table 4b. Series of modified credibility coefficients based on decision rules with confidence values equal to 75% and 100% for set of patients from Table 1.

Conf.		75						100	
Supp.		1		2		3		1	2
Var.		v7		v8		v9		v10	v11
|AR|		15		3		1		14	2
|PR|		14		2		0		14	2
Coeff.		C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}	C_{AR}	C_{PR}
Patients	1	1,0	1,0	1,0	?	1,0	?	1,0	1,0
	2	?	?	?	?	?	?	?	?
	3	1,0	1,0	1,0	1,0	1,0	?	1,0	1,0
	4	1,0	1,0	1,0	1,0	1,0	?	1,0	1,0
	5	?	?	?	?	?	?	?	?
	6	0,3	1,0	0,3	?	0,3	?	1,0	1,0

6 Conclusions

Rough set theory is methodology for knowledge discovery. Credibility coefficients can be applied to identify exceptions to the rules. A more accurate knowledge can be acquired from information system if improper data are removed from it. The main assumption in evaluating credibility coefficients is statement that a majority of data is credible and only a small portion is exceptional. Algorithms of credibility coefficients should identify the both groups.

Objects in decision table can be sorted according to their credibility coefficients. An arbitrary small part of objects with the lowest credibility coefficients can be "suspected to be unusual". They can be removed to improve the quality of the remaining data or can be analyzed with a special attention (to examine exceptions) – both approaches are attractive for a research and can find many reasonable applications.

The methodology of dealing with credibility coefficients requires a lot of further work. New algorithms are being proposed and verified. Only practical results can verify whether credibility coefficients are useful in real applications. We do believe that knowledge includes rules and exceptions and the latter ones should not be neglected.

The idea of assessing, how much an object is typical one in respect to other objects in the set, is a general one. The concept of cataloguing the data by some measures of typicality may be adapted by different data analyzing tools, expert systems, knowledge acquisition systems and many other information processing systems, where detecting of exceptions may be significant or at least useful.

References

1. Podraza R., Walkiewicz M., Dominik A.: *Credibility Coefficients in ARES Rough Set Exploration System,* In: Proc. 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2005, Regina, Canada, 2005, Lecture Notes in Artificial Intelligence, **3642**, Part II, Springer-Verlag Berlin Heidelberg (2005) 29-38.
2. Podraza R., Tomaszewski K.: KTDA: *Emerging Patterns Based Data Analysis System,* Annales Universitatis Mariae Curie-Skłodowska, Sectio AI Informatica, Vol. IV, (2005) 279-290.
3. Podraza R., Dominik A.: *Problem of Data Reliability in Decision Tables,* International Journal of Information Technology and Intelligent Computing , Vol. **1**, No. 1, (2006) 103-112.
4. Podraza R., Walkiewicz M., Dominik A.: *Credibility Coefficients Based on Frequent Sets,* Annales Universitatis Mariae Curie-Skłodowska, Sectio AI Informatica, Vol. V (to appear).
5. Podraza R., Dominik A., Walkiewicz M.: *Application of ARES Rough Set Exploration System for Data Analysi,.* Annales Universitatis Mariae Curie-Skłodowska, Sectio AI Informatica, Vol. III, (2005).
6. Pawlak Z.: *Rough Sets. Theoretical Aspects of Reasoning about Data.* Kluwer, Dordrecht (1991).
7. Pawlak Z.: *Rough Sets,* International Journal of Computer Information Sciences, **11** (1982).
8. Skowron A.: *Extracting Laws from Decision Tables: A Rough Set Approach,* Computational Intelligence, **11** (1995).
9. Podraza R., Dominik A., Walkiewicz M.: *Decision Support System for Medical Applications,* In: Proceedings of the IASTED International Conference on Applied Simulations and Modeling, Marbella, Spain, ACTA Press, Anaheim, Calgary, Zurich (2003) 329-334.
10. Podraza R., Podraza W.: *Rough Set System with Data Elimination,* In: Proceedings of the 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas, Nevada, USA, CSREA Press (2002) 493-499.