Annales UMCS Informatica AI 5 (2006) 69-77

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Credibility coefficients based on frequent sets

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

[1]Institute of Computer Science, [2]Institute of Radioelectronics, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warszawa, Poland

Abstract

Credibility coefficients are heuristic measures applied to objects of information system. Credibility coefficients were introduced to assess similarity of objects in respect to other data in information systems or decision tables. By applying knowledge discovery methods it is possible to gain some rules and dependencies between data. However the knowledge obtained from the data can be corrupted or incomplete due to improper data. Hence identification of these exceptions cannot be overestimated. It is assumed that majority of data is correct and only a minor part may be improper. Credibility coefficients of objects should 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 frequent sets, which are produced while using data analysis based on the rough set theory. Some information on the rough set theory is supplied to enable expression of credibility coefficient formulas. 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 [1-4]were introduced to identify improper objects in information systems or decision tables. Credibility coefficients are defined as a heuristic measure from the range <0.0;1.0>, where the numbers close to the lower bound denote a low credibility, whereas the numbers close to the upper bound denote a high credibility. The whole concept is based on the basic assumption that majority of collected data is trustworthy and only minority of it can be considered as corrupted, improper or exceptional. Based on this assumption some calculations are performed. They are aimed to evaluate similarities between data, which suggest their typicality and hence deduced high credibility. In formulas of credibility coefficients the similarities of attributes are rewarded by increasing a result, while differences are punished by decreasing the output.

[*]Corresponding author: *e-mail address*: R.Podraza@ii.pw.edu.pl

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

Credibility coefficients can identify roughly a set of proper objects and a set of improper ones. The ARES Rough Set Exploration System is a general data analysis tool, but it was designed and developed for medical applications [9,10]. Specifically medicine and other natural sciences are very often oriented toward describing exceptions to the rules especially if the rules are well recognized and accepted. For instance, it is very important to identify a disease when symptoms are misleading, when a case does not fit to the rules. A good physician can be recognized by a way of distinguishing and dealing with exceptions. Credibility coefficients' purpose was to provide an automatic aid in expert systems for identifying such exceptional cases to draw a special attention of specialists to these cases.

The paper comprises a short description of rough set theory to enable presenting the concept and mathematical descriptions of credibility coefficients. In this way a precise and concise presentation of idea of introducing the coefficients can be done. Then follows three chapters presenting respectively algorithm of credibility coefficients evaluated using frequent sets, an example of applying the credibility coefficients and finally a proposal of modification of the algorithm. The paper is completed with some conclusions and suggestions how credibility coefficients can be exploited in practice.

2. Elements of rough set theory

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 $f : U \times Q \to V$ 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 $C \to D$; otherwise is non-deterministic.

Elementary condition is a pair of attribute-value. Every object is represented or satisfy 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 $\langle P \rangle$) 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 $\langle P \rangle$, 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.

3. Algorithm

Descriptions:

W[] – vector *W* whose index domain may be any set of data, in particular for the object $t \in DT$, *W*[*t*] denotes the value of vector element, which is associated with the object *t* (e.g. vectors *counts*[], *decCount*[], *CFS*[]),
t.dec – value of the decision attribute of the object $t \in DT$,
sum(*W*) – sum of all elements of vector *W*,
X.len – length of set *X*,
Inf(*t*) – set of elementary conditions based on values of successive attributes of object *t*.

Algorithm:

Input:
F – set of frequent sets (without the empty set),
DT – decision table,
Output:
CFS[] – vector of credibility coefficient values.

```
1  counts = New counts[]
2  Forall f ∈ F Do
3     decCount = New decCount[]
4     Forall t ∈ DT Do
5        If f ⊂ Inf(t) Then
6           counts[t] := counts[t] + 1
7           decCount[t.dec] := decCount[t.dec] + 1
8     Forall t ∈ DT Do
9        If f ⊂ Inf(t) Then
10          p := decCount[t.dec]/sum(decCount)
11          C_FS'[t] := C_FS'[t] + p * 1/f.len
12 Forall t ∈ DT Do
```

```
13    If (counts[t] <> 0)
14        C_FS[t] := C_FS'[t] / counts[t]
15 normalizeCoeff(C_FS[])
```

A short interpretation of the algorithm (provided below) should explain ideas and motivations in designing the credibility coefficient.

- Vector *counts* (created in line 1) is associated with the number of frequent sets, which are subsets of set of elementary conditions of object *t*. Elements of the vector are modified in line 6 and are used in lines 13-14 for producing values of vector *CFS*.
- For each frequent set (loop in lines 2-11):
  - Vector *decCount* is created (line 3) to keep numbers of objects having the same value of the decision attribute and being supersets for the currently analyzed frequent set. The elements of the vector are updated in line 7.
  - For each object (line 8), which is a superset of the currently analyzed frequent set, its temporary credibility coefficient *C'* is modified (line 11). Value of the credibility coefficient is incremented by a product of reciprocal of the length of the frequent set and a factor *p*. The factor *p* represents ratio of two numbers of objects being supersets of the frequent set. The numerator is a number of objects having the same value of the decision attribute as the considered object and the denominator is a number of all objects which are supersets of the frequents set. Applying reciprocal of the length of the frequent set in the evaluation of the credibility coefficient favours short frequent sets, which are more characteristics of the whole information system. In contradiction, long frequent sets are characteristic of specialized objects.
- The algorithm is completed by averaging all credibility coefficients. Their temporary values ($C_{FS}'[t]$) are divided by number of frequent sets, which were subsets of the particular objects (lines 12-14). The last step (line 15) is scaling performed by function *normalizeCoeff*, presented below.

```
   normalizeCoeff(C[])
1  threshold := 0.9 * max(C[])
2  Forall t ∈ DT Do
3     If (C[t] ≥ threshold)
4        C[t] := 1
5     Else
6        C[t] := C[t] / threshold
```

The important feature of the algorithm of credibility coefficients based on the frequent set is omitting such objects, which have such a typical attribute values that no frequent set (generated with a required support) is a subset of any of them. This fact is indicated by zero value of elements of vector *counts* associated

with these objects. In this version of the algorithm the non-scaled value of the coefficient is set to zero.

Interpretation of the algorithm of function *normalizeCoeff* is as follows. Firstly the maximum value of all coefficients is found (applying function *max*) and a threshold is set to 90% of this value. All objects with the credibility coefficient higher than the threshold are considered as perfectly credible or typical and their credibility coefficients are updated to 1. This reflects the assumption that a deviation up to 10% from the "best" object is negligible and entitles the object to be "perfect". All other coefficients are modified by dividing their values by the threshold. In this way the values from a narrower interval are extended to the domain presumed for credibility coefficients (interval <0.0;1.0>.

More formally a credibility coefficient C_{FS} for object $u \in U$ of a decision table $DT = (U, C \cup \{d\}, V, \varphi)$ can be expressed as:

$$C_{FS}(u) = \begin{cases} 1 & \text{for } C'_{FS}(u) \geq 0.9 \cdot \max\left(C'_{FS}\right) \\ C'_{FS}(u) \cdot \dfrac{1}{\max\left(C'_{FS}\right)} & \text{otherwise} \end{cases}$$

where:

$$C'_{FS}(u) = \frac{\sum_{f \in F} \frac{1}{|f|} \cdot \frac{\left|\left\{t \in U : t \in \langle f \rangle \wedge f(u,d) = f(t,d)\right\}\right|}{\left|\langle f \rangle\right|}}{\left|\left\{f \in F : u \in \langle f \rangle\right\}\right|},$$

$|f|$ – length of set f (number of its elementary conditions),

$\max(C'_{FS})$ – element with maximum value from set $\left\{C'_{FS}(i) : i = 1 \ldots |U|\right\}$.

4. Example

Application of the credibility coefficient based on frequent sets is presented in the 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 the form of texts (representing values of enumerations) and the corresponding integer number (coded data in parentheses). The decision table is extended by a column with the values of credibility coefficients based on frequent sets with the minimum support set to 40% (at least two objects).

The credibility coefficients different from 1.0 were evaluated for the objects with numbers 2, 5 and 6. Objects 2 and 5 are incredible, because they introduce indeterminism in the decision table. Let us consider credibility coefficients generated for frequent sets with different values of minimal support. The results are presented in Table 2.

Table 1. Decision table with credibility coefficients based on frequent sets with minimum support of 40%

Patient	Headache (g)	Myalgia (m)	Temperature (t)	Flue (f)	C_{FS}
1	No (0)	Yes (1)	High (0)	Yes (1)	1.00
2	Yes (1)	No (0)	High (0)	Yes (1)	0.83
3	Yes (1)	Yes (1)	Very High (1)	Yes (1)	1.00
4	No (0)	Yes (1)	Very High (1)	Yes (1)	1.00
5	Yes (1)	No (0)	High (0)	No (0)	0.67
6	No (0)	Yes (1)	Normal (2)	No (0)	0.76

Table 2. Values of credibility coefficients based on frequent sets with different values of minimal support

		Minimal Support [in number of objects]			
		1	2	3	4
Patient	1	1.00	1.00	0.93	0.97
	2	0.90	0.83	1.00	1.00
	3	1.00	1.00	1.00	0.97
	4	1.00	1.00	0.93	0.97
	5	0.81	0.67	0.48	0.00
	6	0.95	0.76	0.38	0.28

Table 3. Values of credibility coefficients based on frequent sets without Object 5

		Minimal Support [in number of objects]			
		1	2	3	4
Patient	1	1.00	0.86	0.72	0.97
	2	1.00	1.00	1.00	1.00
	3	1.00	0.94	0.83	0.97
	4	1.00	0.81	0.72	0.97
	6	0.88	0.35	0.27	0.28

For value 1 of minimal support the only object with credibility coefficient below 0.9 is object 5. For all columns objects 1, 3 and 4 have values above 0.9. For object 2 the credibility coefficients are at least above 0.8. The credibility coefficients rapidly decrease for higher values of the minimal support of frequent sets, because there are too few objects with decision of objects 5 and 6 ("Flue=No").

From Table 2, it can be observed that the least credible object is object 5. We treat its data as improper. Let us see consequences of removing it from the decision table.

All credibility coefficients for objects 1, 2, 3 and 4 have values above 0.7. Credibility coefficient values of object 6 decrement with increasing values of

minimal support of frequent sets. High value of credibility coefficient for the minimal support of 1 can be explained that only this value of support enables generating frequent sets with elementary conditions typical only of this object ("temperature= Normal" and "flue=No").This is an evidence that credibility coefficients based on frequent sets with relatively high support may identify objects, which can be just rare and hence require more attention from the expert, if we are interested in non-typical data.

5. Modification of credibility coefficient

The credibility coefficients based on frequent sets have one drawback. They poorly deal with objects that do not match frequent sets. This situation is caused by a limited number of frequent sets, which is a consequence of a value of minimal support. Considering such objects, which are being named uncertain, the credibility coefficient based on frequent sets is fixed to the minimal value (zero). This problem can be handled in a number of ways:

- supply an excessive set of frequent sets (in particular with minimal support set to 1),
- set for such objects and arbitrary value,
- extend the domain of credibility coefficients by introducing an extra denotation for uncertain objects.

The first two approaches are difficult to be implemented in a general approach. Excessive frequent sets may lead to unacceptable processing time, while choosing an arbitrary value form the domain of credibility coefficient may cause ambiguities in interpretation of results. According to the assumption expressed in point c. a new value, namely -1, is introduced to the domain of credibility coefficient to denote an uncertain object.

To modify the algorithm presented in chapter 3 it is enough to add at the end of the algorithm a loop presented below. In the loop all credibility coefficients with value set to zero are set to the value -1 denoting an uncertain object. The modified credibility coefficient is represented as C_{FS}^{M})

```
1-15  ...
16    Forall t ∈ DT Do
17        If (counts[t] = 0)
18            C^M_FS[t] := -1
```

The formal description of the modified credibility coefficient C_{FS}^{M} for the object $u \in U$ of a decision table $DT = (U, C \cup \{d\}, V, \varphi)$ can be presented as:

$$C_{FS}^{M}(u)=\begin{cases}-1 & \text{for } C_{FS}^{'}(u)=0\\ 1 & \text{for } C_{FS}^{'}(u)\geq 0.9\cdot\max\left(C_{FS}^{'}\right)\\ C_{FS}^{'}(u)\cdot\left[1/\max\left(C_{FS}^{'}\right)\right] & \text{otherwise}\end{cases}$$

where:

$$C_{FS}^{'}(u)=\frac{\sum_{f\in F}\frac{1}{|f|}\cdot\frac{\left|\left\{t\in U: t\in\langle f\rangle\wedge f(u,d)=f(t,d)\right\}\right|}{\left|\langle f\rangle\right|}}{\left|\left\{f\in F: u\in\langle f\rangle\right\}\right|},$$

$|f|$ – length of set f (number of its elementary conditions),

$\max(C_{FS}^{'})$ –element with maximum value from set $\left\{C_{FS}^{'}(i): i=1...|U|\right\}$.

$C_{FS}^{M}(u)=-1$ denotes an uncertain object.

The introduced modification is important, because the objects, which cannot be properly identified by the modified credibility coefficients based on frequent sets, get a special denotation. This extra value can be understood as exception and an interpretation is a task of an expert. Anyway, uncertain object do not contribute much to the knowledge induced from the decision table, and treatment of such data depends on a purpose of applying knowledge discovery techniques, vulnerability of data or expert approach.

Conclusions

Rough set theory provides methodology for automatic knowledge acquisition. The methodology can be refined by applying credibility coefficients to identify exceptions to the rules. More precise classification (with better quality indicators) can be obtained from an information system if improper data is removed from it. Analysis of exceptions can very often enhance quality of data collecting, processing and storing (by reducing errors).

Objects in the decision table can be sorted according to their credibility coefficients. A 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 care (to observe an exception) – both approaches are interesting for research and can find many reasonable applications.

In interpretation of credibility coefficients it has to be assumed that majority of data are credible and only small portion is exceptional data. Heuristic algorithms of credibility coefficients should reveal similarities of groups of objects. Then a typical objects can be pointed out as not belonging to these groups.

The methodology of dealing with credibility coefficients requires a lot of efforts to be developed. New algorithms for credibility coefficients are being proposed and verified. And only the practice can prove whether credibility coefficients will supplement expert systems. We do believe that knowledge consists of two parts: rules and exceptions and the latter one should not be neglected.

The idea of assessing, how much one object is typical in respect to other objects in the set, is a general one. The concept of weighting the data by some measures of typicality (recognized by frequency of appearing) may be adopted by different data analyzing tools, expert systems, knowledge acquisition systems and many other information processing systems, where detecting of exceptions may be important.

References

[1] Podraza R., Walkiewicz M., Dominik A., *Credibility coefficients in ARES rough set exploration system*, Proc. 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2005, Regina, Canada, August/September, 2005, Lecture Notes in Artificial Intelligence, LNAI 3642, Part II, Springer-Verlag, (2005) 29.

[2] Podraza R., Tomaszewski K., *KTDA: Emerging patterns based data analysis system*, XXI Fall Meeting of Polish Information Processing Society, Conference Proceedings, Wisła Poland, 2005 213, on CD-ROM.

[3] Podraza R., Dominik A., *Credibility coefficients for objects of rough sets*, Proceedings of VII International Conference on Artificial Intelligence AI-20'2005, Artificial Intelligence Studies, Special Issue, 2(25) (2005) 205.

[4] Podraza R., Jurkowski A., *Coefficient of credibility in rough set system*, Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria, ACTA Press, (2004) 776.

[5] Podraza R., Dominik A., Walkiewicz M., *Application of ARES Rough Set Exploration System for Data Analysis*, Conference Computer Science – Research and Applications, Kazimierz Dolny, Poland, Annales Universitatis Mariae Curie-Skłodowska, Sectio AI Informatica, II (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, (1982) 11.

[8] Skowron A., *Extracting Laws from Decision Tables: A Rough Set Approach*, Computational Intelligence, (1995) 11.

[9] Podraza R., Dominik A., Walkiewicz M., *Decision Support System for Medical Applications*, Proceedings of the IASTED International Conference on Applied Simulations and Modeling, Marbella, Spain, (2003) 329.

[10] Podraza R., Podraza W., *Rough Set System with Data Elimination*, Proceedings of the 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'2002), Las Vegas, Nevada, USA, (2002).