Annales UMCS Informatica AI 2 (2004) 163-172

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

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

Ignoring improper data in decision support system for medical applications[+]

Roman Podraza[a*], Piotr Ryszkowski[a], Wojciech Podraza[b]

[a]*Institute of Computer Science, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warszawa, Poland*
[b]*Department of Medical Physics, Pomeranian Medical University, Szczecin, Poland*

Abstract

A Decision Support System for Medical Applications was designed by applying the rough set theory to generate rules from the collected data. The data are kept in a table representing information system. There are some improper data in information systems and their removal can improve the quality of the retrieved information. By improper data we can understand such objects that disturb rules generation. They can be erroneous or corrupted or just exceptions. It is possible to find an algorithm of improper data removal to optimize the quality of information derived from decision tables. The improper data can be verified by checking whether some indicators of classification quality were improved after removal of the data. Some suggestions of identifying improper data are presented in the paper. In medical applications the improper data cannot be neglected.

1. Introduction

In medicine and other natural sciences there is still a lack of scientific methodologies for extracting knowledge, representing intuition on observed reality, finding relations between data, deriving suppositions. The rough set theory [1,2] is a good formal tool that can be used in developing such an approach. The decision tables [3-5] known from the rough set theory are recognized and accepted tools for deriving rules from data. In the paper an outline of the Decision Support System for Medical Applications is presented. The system applies the rough set theory as an engine for deriving rules.

The system is a general one and can be applied to any problem. Elimination of improper data [6,7] was introduced as a new property. By the improper data we assume such elements, which disturb a process of deriving rules from other data. A simple but convincing example of improper data and their impact on

[+] After second revision.
[*] Corresponding author: *e-mail address*: rpd@ii.pw.edu.pl

decision table is presented in the paper. In rough set theory some quantitative indicators represent information on data coherence and data accuracy. The indicators like accuracy of approximation and quality of approximation denote the quality of rules, facts and relationships derived from the data tables.

The improper data can be properly identified if elimination of relatively "small" portion of objects results in improving (maximizing) the quantitative indicators. There is a maximum fraction of objects from decision table that can be considered as improper. It is decided by an expert. The improper data should be recognized, but the reality should not be just enhanced. The approach is perhaps important in most applications, but is essential in medicine [8]. There are very often interactions of different diseases, various treatments, reactions to drugs and so on. Such cases can blur the general rules. In typical rough set approach it is usually suggested to increase a number of attributes to discriminate the data better, to find more precise solution to the ambiguous cases. Sometimes, maybe, it can be prepared for a repeatable process. In medicine it is usually impossible or very hard to add extra information to history of a treatment [9]. On the other hand, it is very interesting and very important for the treatment to find the special cases potentially denoted by improper data. Medical applications should be sensitive to improper data while industrial application mostly can neglect such a data.

In the Decision Support System for Medical Applications the improper data are removed "gently". They are marked as improper and are not taken into consideration for rough set analysis (and then we should observe better values of the quantitative indicators). The user can mark a chosen object arbitrarily as improper one or remove the marking from the data previously designated by the system or by the user. The facility of manipulation of the improper data marking seems to be useful in tuning medical experiments on drawing conclusions from the data set. The improper data can be injected to test different algorithms of their recognition.

2. Rough Set Theory

Rough set theory is used for analyzing data in an information system. The information system S can be defined as

$$S = < U, Q, V, \rho > ,$$

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 $\rho : U \times Q \to V$ is a function that $\rho(x, q) \in V_q$ 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. Values of function ρ are shown in the table cells (see Table 1).

Let $S =< U,Q,V,\rho >$ be an information system and $P \subseteq Q$, and $x,y \in U$. Objects x and y are indiscernible by set of attributes P (denoted by x $\tilde{P}$ y) in S iff $\rho(x, q) = \rho(y, q)$ for every $q \in P$. The indiscernibility relation $\tilde{P}$ is an equivalence relation on the set of objects U.

Let P^* denote family of all equivalence classes of relation $\tilde{P}$ on U. Equivalence classes of $\tilde{P}$ on U are called P-elementary sets in the information system S. $Des_P(X)$ denotes a description of P-elementary set $X \in P^*$.

$$Dec_P(X) = \{(q,v) : \rho(x,q) = v, \text{ for all } x \in X \text{ and all } q \in P\}.$$

For any set $Y \subseteq U$ and attributes $P \subseteq Q$ it is possible to define P-lower approximation of Y in the information system S as

$$\underline{PY} = \bigcup_{X \in P^* \wedge X \subseteq Y} X$$

and P-upper approximation of set Y in the information system S as

$$\overline{PY} = \bigcup_{X \in P^* \wedge X \cap Y \neq \emptyset} X .$$

The P-boundary of Y is defined as

$$Bn_P = \overline{PY} - \underline{PY} .$$

The accuracy of approximation of set Y by set of attributes P in the information system S can be defined as

$$\mu_P(Y) = \frac{card(\underline{PY})}{card(\overline{PY})},$$

where *card* is cardinality of the set.

Let $P \subseteq Q$ be a set of attributes and $Y = \{Y_1, Y_2, \ldots Y_n\}$ be family of sets

where $Y_i \cap Y_j = \emptyset$ for all $i, j \leq n$

and $\bigcup_{i=1}^{n} Y_i = U$

P-lower and P-upper approximations of family of sets Y in the information system S are respectively the sets

$$\underline{PY} = \{\underline{PY_1}, \underline{PY_2}, ..., \underline{PY_n}\}$$
$$\overline{PY} = \{\overline{PY_1}, \overline{PY_2}, ..., \overline{PY_n}\}$$

The quality of approximation of partitioning of Y by a set of attributes $P \subseteq Q$ is

$$\chi_P(Y) = \frac{\sum_{i=1}^{n} card(\underline{PY_i})}{card(U)}.$$

An information system can be regarded as a decision table if the set of all attributes is split into condition attributes C and decision attributes D

$$Q = C \cup D \text{ and } C \cap D = \varnothing.$$

The information system $S = <U, C \cup D, V, \rho>$ is deterministic iff $C \rightarrow D$; otherwise it is non-deterministic.

Let $C^* = \{X_1, X_2, \ldots X_k\}$ and $D^* = \{Y_1, Y_2, \ldots Y_n\}$. A decision rule in information system S is denoted as

$$Des_C(X_i) \Rightarrow Des_D(Y_j).$$

The set of decision rules $\{r_{i,j}\}$ for every class Y_j is defined

$$\{r_{i,j}\} = \{Des_C(X_i) \Rightarrow Des_D(Y_j),\ X_i \cap Y_j \neq \varnothing,\ i = \{1,2,\ldots,k\},\ j = \{1,2,\ldots,n\}\}.$$

Rule $r_{i,j}$ is deterministic iff $X_i \cap Y_j = X_i$, otherwise it is non-deterministic.

3. Elimination of Objects

A decision table is non-deterministic if in an elementary set defined by the conditional attributes there are objects belonging to more than one category defined by the decision attributes. It can be informally stated that data are non-deterministic (imprecise) if indiscriminate objects belong to two or more different sets. Imprecise data can result from insufficient recognition and adding extra attributes can resolve the ambiguities. Anyway such data are potentially improper.

Some part of collected data can be corrupted; some of them can represent exceptions to the rules. Proper identification of the data can enhance the quality of derived rules.

Let us introduce a threshold for removal of improper data, to keep modification of information system under reasonable constraints. **Improper Data Total Threshold** (IDTT) is a number of total data that can be ignored from the whole decision table (usually IDTT is expressed as a percentage relative to the total number of objects). More flexibility in disregarding improper data can be achieved by using the second threshold namely **Improper Data Elementary Threshold** (IDET), which denotes a percentage of objects that can be removed from each elementary set characterized by the conditional attributes. Both limits imposed by IDTT and IDET have to be met.

Let us consider an example of information system presented in Table 1.

In the decision table presented in Table 1 we have three conditional attributes: P = {A, B, C} and one decision attribute {D}. The conditional attributes divide all objects into two elementary sets

$$X1 = \left\{ x \in U \mid_{\{A=1,\ B=1,\ C=1\}} \right\} = \{x_1, ..., x_{50}\}$$

and

$$X2 = \left\{ x \in U \mid_{\{A=2,B=2,C=2\}} \right\} = \{ x_{51}, ..., x_{100}\}$$

Table 1. Non-deterministic Decision Table

X	A	B	C	D
1	1	1	1	1
…	1	1	1	1
23	1	1	1	1
24	1	1	1	2
…	1	1	1	2
48	1	1	1	2
49	1	1	1	0
50	1	1	1	0
51	2	2	2	0
…	2	2	2	0
98	2	2	2	0
99	2	2	2	1
100	2	2	2	2

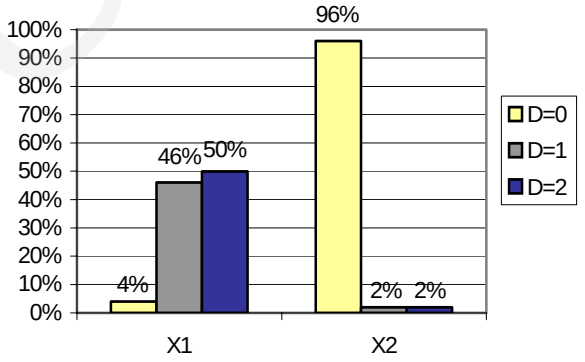

Fig. 1. Distribution of objects in atoms of elementary sets X1 and X2

Decision attribute D has a domain consisting of three values $V_D = \{0, 1, 2\}$. Figure 1 presents how decision categories appear in the elementary sets X1 and X2. Each such category we are going to call atom – all objects belonging to an atom are exactly the same.

Let us have both values of IDTT (a threshold for all objects) and IDET (a threshold for each elementary set) equal to 3%. We can ignore up 3 objects from the whole table (because of IDTT) and only by one object from each elementary

set. The restrictions on elementary sets are more limiting. To have qualitative modification in a set of rules derived from the decision array it is necessary to eliminate whole atoms of an elementary set. It is impossible to remove any atom from the elementary set X1. Minimum quantity of atoms in X1 is two, which is more than the value of IDET defining maximum amount of objects that can be ignored. For the elementary set a number of objects in atoms is 48, 1 and 1 respectively. It is possible to remove one of the atoms, but still elementary set X2 will be classified imprecisely (by 2 different atoms).

If we choose IDTT equal to 3% and IDET set to 5% then we can remove from the elementary set X1 the atom with decision D=0 (objects x_{49} and x_{50}) and then from the elementary set X2 the atom with decision D=1 (object x_{99}) or the atom with decision D=2 (object x_{100}). A distribution of objects in the elementary set after removing objects x_{49}, x_{50} and x_{99} is presented in Figure 2.

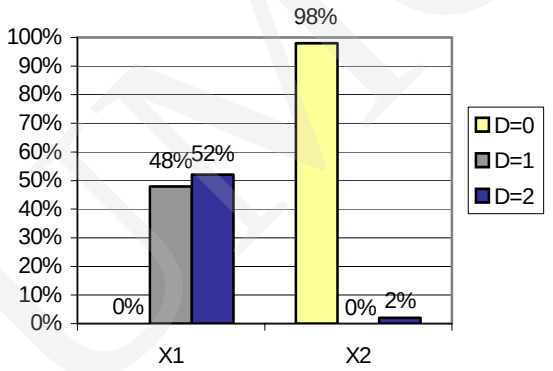

Fig. 2. Distribution of objects in atoms of X1 and X2 after the objects x_{49}, x_{50} and x_{99} were eliminated

The result is not fully satisfactory although we managed to remove atoms from both elementary sets. Anyway, the elementary sets remain still imprecise. If we begin removing objects from the elementary set X2 we can abandon the objects x_{99} and x_{100} (each object is an atom of elementary set X2). Then the value of IDTT prevents from eliminating atoms from the elementary set X1. The results of the operations are shown in Figure 3. In this case by the elimination of only 2% of all objects the elementary set X2 is classified precisely.

For the thresholds of improper data set to some values it is important in which sequence the elimination of objects (potentially improper) takes place. After ignoring improper data, the quantitative indicators are evaluated and the best solution is accepted.

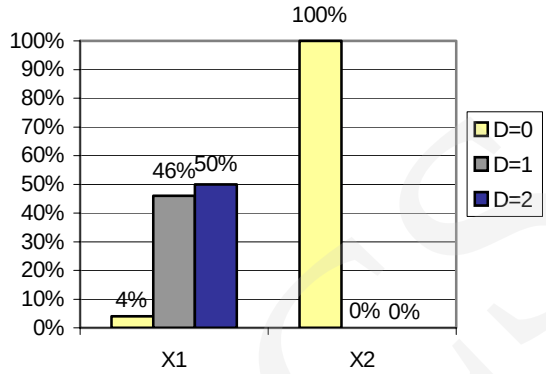

Fig. 3. Distribution of objects in atoms of X1 and X2 after the objects x_{99} and x_{100} were eliminated

4. Algorithm of identifying improper data

Below the algorithm of identifying improper data is presented. Some parts of it typical of generation rules with support of rough set theory are omitted, but the idea of recognizing improper data is publicized. All data names are explained and function names are expected to be self-explanatory.

```
Algorithm: Identifying improper data
Parameters
   Information System SI = (U, Q, V, f)
   C – set of conditional attributes
   D – set of decision attributes
     C⊂Q, D⊂Q, and C∩D=∅ and C∪D=Q
   A –atom – set of objects indiscernible by attributes
    from set C∪D
   A = record
    Objects – list of objects belonging to the atom
    Count – number of objects in the atom
   end
   X – elementary set – set of objects indiscernible by
    attributes from set C
   X = record
    Atoms – set of atoms belonging to the elementary set
    Count – number of atoms in the elementary set
    Inequality – coefficient determining inequality in
             atom distribution in the elementary set
    Idet – threshold of improper objects in elementary set
             – number of objects that can be removed from
             the elementary set
   end
   U – universum - family of all elementary sets Xi
   U = record
```

```
     ElementarySets – family of all elementary sets X_i
     Count – number of elementary sets
     Idet – threshold of improper objects in universum –
              number of objects that can be removed from
              the universum
   end
Results
   Eliminated – set of atoms containing objects identified
            as improper
   Eliminated.Cardinality – number of improper objects; a
            sum of objects in atoms belonging to Eliminated

Procedure
  Candidates – set of atoms selected for elimination
  Candidates.Cardinality – number of objects in all atoms
      selected for elimination
begin
U.ElementarySets := EvaluateElementarySets(C)
for i = 1 to U.Count
   begin
  U.ElementarySets(i).Atoms :=
   EvaluateAtomSet(U.ElementarySets(i),D)
  U.ElementarySets(i).Inequality :=
   EvaluateInequality(U.ElementarySets(i))
   end
U.SortElementarySetsByInequality
for i = 1 to U.Count
   begin
  Candidates :=
   FindEliminationCandidates(U.ElementarySets(i))
  if (Eliminated.Cardinality + Candidates.Cardinality) >
         U.Idtt then
    exit
      else
    Eliminated.Add(Candidates)
   end
end

Function FindEliminationCandidates(X)
Candidates – set of atoms selected for elimination
Candidates.Cardinality – number of objects in all atoms
     selected for elimination
begin
Candidates := ∅
if X.Count > 1 then
   begin
   X.SortAtomsByCountAscending
```

```
   for i = 1 to X.Count
     if (Candidates.Cardinality + X.Atoms(i).Count) >
   X.Idet then
        return Candidates
        else
        Candidates.Add(X.Atoms(i))
   end
return Candidates
end
```

Line 4 of the procedure of identifying improper data represents evaluation of a coefficient denoting a level of inequality of distribution of objects in an elementary set. The result of function `EvaluateInequality` enables assessment whether elimination of objects from the elementary set can improve its properties. Various versions of the function can be used depending on the measure applied. The following statistic indicators of data distribution can be used: entropy, Gini's coefficient or Herfindahl's coefficient. In [10] evaluation of coefficient of credibility was proposed to denote level of influence of every object onto the rules derived from a decision table. In line 5 of the procedure of identifying improper data and in line 3 of the function `FindEliminationCandidates` sorting of objects in the elementary set is performed in such a way that removing a limited number of objects should result in the best improvement. The result should be verified by analyzing the new information system and comparing it with the original one.

5. Conclusions

Rough set theory provides a proper methodology for automatic knowledge acquisition. The methodology can be further refined by applying removal of improper data.

The removal of improper data can reveal dependencies between the other data and generate valuable and important rules. To identify improper data different statistic measures are going to be applied in heuristic algorithms.

In the Decision Support System for Medical Applications the algorithm of identifying and ignoring improper data has been implemented to generate more (and/or better) rules. The improper data are not neglected in the medical applications.

Acknowledgment

The work presented in the paper has been partially supported by the Polish State Committee for Scientific Research under Grant 7 T11C 012 20.

References

[1] Pawlak Z., *Rough Classification*, International Journal of Man-Machine Studies, (1984) 20.
[2] Pawlak Z., *Rough Sets*, International Journal of Computer Information Sciences, (1982) 11.
[3] Boryczka M., Słowiński R., *Derivation of Optimal Decision Algorithms from Decision Tables Using Rough Sets*, Bulletin of the Polish Academy of Sciences, Technical Sciences, (1988) 36.
[4] Pawlak Z., *Decision Tables and Decision Algorithms*, Bulletin of the Polish Academy of Sciences, Technical Sciences, (1985) 33.
[5] Skowron A., *Extracting Laws from Decision Tables: A Rough Set Approach*, Computational Intelligence, (1995) 11.
[6] Podraza W., *The Method of Reduction of Rough Sets Theory Results Misinterpretation In Medicine*, Proceedings of Second Symposium on Modelling and Measurements in Medicine, Krynica, Poland, (2000), in Polish.
[7] 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).
[8] Yasdi R., *Combining Rough Set Learning- and Neural Learning-method to Deal with Uncertain and Imprecise Information*, Neurocomputing, (1993) 7.
[9] Stefanowski J., Słowiński K., *Rough Sets as a Tool for Studying Attribute Dependencies in the Urinary Stones Treatment Data Set*, T. Y. Lin, N. Cercone eds.: Rough Sets and Data Mining. Analysis for Imprecise Data, Kluwer Academic Publishers, Boston/London/ Dordrecht, (1997).
[10] Podraza R., Jurkowski A., *Coefficient of Credibility in Rough Set System*, Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA'2004), Innsbruck, Austria, (2004).