UDC 004.05

Shakhovska N.
Information Systems and Networks Department,
Lviv Polytechnic National University,
S. Bandery Str., 12, Lviv, 79046, Ukraine,
E-mail: natalia_s@rambler.ru

THE MODELLING UNCERTAINTY IN DATAWAREHOUSES, BASED ON RELATIONAL MODEL

In this paper the methods of construction of classification rules for elimination of equivocations are described. The algorithms for the solution of primary goals are offered.

Keywords – classification rule, datawarehouse, tuple, uncertain data.

1. Introduction

In quantities of data domains it is necessary to process the indistinct information, at what the outcome of a data analysis completely depends on a degree of their entirety in a system. The representative data domains of appearance of an illegibility are a sociological orb (exchange of activity, public funds, marketing researches of the market etc.), historical researches, planning of economic activities etc.

In the article the algorithms of classification and classifying of objects are tendered, the information on which one is saved in repository.

The main problems, which one arise in problems to the analysis and structuring of the data, are the problems of creation of classes and reference to them of objects, the information on which one just has entered the database. The problem of creation of the class contain two subtasks:

1) Construction of classification functions, according to which one the object is categorized as the quote of the definite class;
2) Breakdown on classes and identification of the obtained classes.

In the article we shall consider maiden from these subtasks.

2. Review of modern researches and allocation of unsolved problems

At a level of a tuple in repository is entered 8 types of equivocations [1-3].

1. Value is unknown (missing).
2. Incompletness of the information
3. Illegibility (usage of distribution for installation of the variety of knowledge)
4. The inaccuracy (concerns numerical data)
5. Non-determination of conclusion procedures of the solutions
6. Unreliability of the data
7. Multivalence of interpretations
8. Linguistic indefinability.

Let us consider the more detailed indicated types of equivocations and find out places of their occurrence in relation.

Uncertainty of types 3-8 categorize in [1] as wobble of the data and predominantly occur at a level of a tuple or subset of values of attributes.

The zero information most often meets at a level of attribute value. The incompletness is a condition of a tuple, in which there are missing values. It is possible to attribute an illegibility, inaccuracy and contingency to physical uncertainty, one of sources at which one is limitation exactly of numeric data types or loss of accuracy in a run time of mathematical operations (here attribute uncertainty arises owing to activity with intervals). The unreliability and multivalence of interpretations arises in connection with inexact analysis or ambiguous mapping of objects in relation. In relation is figured with the help of padding attribute, the characterizes values which measure of confidence to a tuple or subset of values of attributes in a tuple. The multivalence of interpretation is by one of sources of originating of inconsistencies. The linguistic uncertainty is connected with usage of natural language for knowledge submission, which haves qualitative nature, and there can be owing to misunderstanding value of a word or misunderstanding of the contents the proposal. Such type of uncertainty meet in systems of text information processing (machine translation system, self-conditioning system etc.).

The reviewed types of equivocations can be superimposed against each other or to be a source of occurrence one another.

Nowadays time the methods of elimination are missing, inexact and indistinct data [1-3] are designed. Therefore it is necessary to elaborate methods, which can work with all types of uncertainty.

3. Formulation

Let us we have relation *r* with the scheme R. We must to construct a set of classification rules s ($X \rightarrow Y$), where $X, Y \subset R$, $X \cap Y = \varnothing$; X - the subset of attributes, on the basis of values which implements reference to the class (elimination of uncertainty on values of attribute Y), Y - attribute (subset of attributes).

Input data for classification (the reference to the class) is a range of *target* attributes. Target attributes(X) are used for a data analysis, and pursuant to values which breakdown on classes implements. To target attributes we consist all attributes, which one enter in set of keys. To target attributes we shall relate all attributes, which enter in a set of the left-hand parts of functional connections (except for primary keys), and those attributes, which will influence a degree of confidence to the obtained outcome of the analysis. Besides for a concrete data domain with the help of expert interrogation the padding subset of attributes is determined, which be considered us target for the analysis. For example, for a problem of sociological interrogation such attributes are age, education, payment etc.

The attributes, above which operations of an aggregate and matching are executed, we shall call *critical* Y (submit outcomes of the analysis). Critical attributes contain numerical data, uncertainty, introduced in view, and right members of functional relations. Let us also concern to critical attributes, which one contain titles of classes (label) [4].

Class is a subset of tuples for which value on set of critical attributes is identical

$$cl = \sigma_r (X = x, Y = y).$$

To simplify of a problem we shall consider, that the classes are definite, and their characteristics (that is title and rule, behind which the object is considered as the quote of this class) are saved in the database.

The reference to the class implements on the basis of definition of a subset of values of target attributes. For example, for the “Student” class the value of target attributes should be contented with conditions: age - [16, 23], education - {mean, mean professional, unfinished maximum}, payment - [150 $, 250 $].

In connection with that is high-gravity to receive the full information on objects of a data domain, there not all target attributes can be definited. Therefore for each class the value of limit is determined, which one means a minimum degree of confidence to object, behind which the object can be categorized as the quote of this class. The degree of confidence s to object is determined as relation of target attributes with defined values to all definite target attributes of this class (the greater it is known about object, the maximum will be a degree of confidence).

$$s = \sum \begin{cases} 0, cr_i \notin \text{cl} \\ 1, \ cr_i \in \text{cl} \end{cases},$$

Where cr_i - value on set of target attributes.

Let us consider that if value of attribute is definite, it is authentic.

Let us consider the problem eliminations of uncertainty.

The reference to the class can be esteemed as one of ways of elimination of uncertainty, you see in process classification the filling of empty value of attribute implements, which one contains value of a title of the class. Besides it is possible to consider classification rules as indistinct functional connections relations.

In databases the indistinct functional connection is supported:

$$\text{E} (\text{X} \rightarrow \text{A}).$$

If the ratio of tuples, on which this functional connection is executed, to tuples, on which she is defaulted, not smaller, than s, where s - value of limit the miss, definite on the basis of expert interrogation [3]. Certainly, value s - not smaller value of a class limit.

$$e(X \to Y): \frac{COUNT(X = x, Y = y)}{COUNT(X = x, Y \neq y)} \geq s_{cl}$$

Value of limit of the miss we shall mean by a degree the multivalued logics of Lukasiewicz (changes in borders(limits) [0, 1]).

From here follows, that the algorithms of elimination of equivocations with the help of functional connections can be applied for objects classification.

Let us have for example, quality of an example, in the database there is a classification rule Education, Payment, Age → Social group.

4. Base material

To have a capability to categorize objects, it is necessary to construct functions of classification. In general, in the database the information on several types of classes can be saved, and for each type of the class there is a subset of functions. The same function can be applied to definition of several types of classes.

Let us consider algorithm of spawning of classification functions (rules).

The rules can be generated by two ways:

- On the basis of the analysis of the characteristics of classes;
- On the basis of the existing rules.

Spawning classification rules on the basis of the analysis of the characteristics of the class

In case of application of the choice way, the classification rules, first of all, will be plotted on the basis of functional connections, which are supported in relation. The degree of confidence to such rule will be maximum (And = 1).

Subset of attributes, which will go into rules, are determined on the basis of the analysis of the characteristics class.

Sequence of steps:

1. The tuples of relation are assorted behind titles of classes.
2. Inside group passes in turn grouping behind each target attribute.
3. If quantity of members of a subgroup, switching on with empty values, does not equal quantities of tuples in group of the class, other attribute for check is elected and is transferred on a step 2.
4. Is spotted values *e* as relation of quantity of tuples with nonblank value parsed to attribute to quantity of all tuples in group (that is spotted response frequency curve).
5. To the obtained response frequency curves is used multivalued “or”: $u \;\&\; v = \max\{0, u + v - 1\}$,
6. Classification rules as the left-hand part all attributes will include, the frequency response curves which more or less to value obtained on a step 6, and the frequency response curve will be considered as a degree of confidence to the rule.

Construction of classification rules by a sweep method

Let us consider one of ways of elimination of a critical values. Classification rule we shall consider as an approximated functional connection with a definite degree of confidence We use for this purpose a method similar to a known sweep method [2]: the equalling of values of attributes in the left-hand part of a rule with a degree of confidence And means also equalling of values of attributes in a right member.

Let us describe algorithm of application of a modified sweep method.

Let in relation r the approximated functional connection is supported

e (X1..., Xn→ A). A character ↓ means a defined value, and ⊥ - its absence; t_i - tuple of relation *r* (sequence of tuples has no meaning)

1. If {t1 (X1)↓..., t1 (Xn)↓} and {t2 (X1)..., t2 (Xn)↓, and {t1 (X1)↓..., t1 (Xn)↓ =
    i. t2 (X1)↓..., t2 (Xn)↓} and {t1 (A)↓} and {t2 (A) = ⊥}, is changed at each entering ⊥ in *r* on t1 (A).
2. If {t1 (X1)↓..., t1 (Xn)↓} and {in t2 m with n of values of attributes -↓, n - m of values of attributes - ⊥, $m \leq n$ } and { $e \leq m/n$ } And {on defined values t1 (Xm)↓ = t2 (Xm)↓} and {t1 (A)↓} and {t2 (A) = ⊥} and is changed at each entering ⊥ in *r* on t1 (A).
3. If {in ti mi with n of values of attributes -↓, $m_i \leq n$ } and {in tj mj with n of values of attributes -↓, $m_j \leq n$ } and {on defined values ti (Xm)↓ = t2 (Xm)↓} and {on defined values tj (Xm)↓ = t2 (Xm)↓} and { $m_i/n \leq m_j/n$ } And {ti (A)↓} and {tj (A)↓} and {t2 (A) = ⊥}, is changed at each entering ⊥ in *r* on tj (A).

Spawning classification rules on the basis of existing rules

In [2] usage of degrees the multivalued logicians for submission of confidence to the rule is demonstrated. For such submissions of dextral and left-hand parts of a rule it is possible to consider(count) discrete, and to work with them routined as with separate parts. As in precursor section is rotined, that the classification rule is considered as an approximated functional connection, it is possible to them to apply the main axioms of a conclusion [3]. We can to use logic operations of the multivalued logicians [1] “and” for child’s and “or” for the ancestors, we receive a capability to generate new rules on the basis of existing and automatically to determine to them degrees of confidence (which one can be tested experimentally). From here follows, that it is necessary in the database to save only minimum cover of approximated functional connections (that is classification rules), and the remaining rules can be added on the basis of their speed keys with usage of operations the multivalued logics [1] and axioms of a conclusion.

Example of spawning of the rules:

TABLE 1

Example of generating of the rules on the basis of existing rules

Existing rules	**Born rules**
Age, Education $\xrightarrow{0,8}$ Social group Payment $\xrightarrow{0,4}$ Social group	Age, Education, Payment $\xrightarrow{0,4}$ Social group
Age, Education $\xrightarrow{0,8}$ Social group Education $\xrightarrow{0,4}$ Level of supply of materials	Age, Education $\xrightarrow{0,2}$ Level of supply of materials, Social group

The classification rules (or indistinct functional connections) are expedient for saving in separate relation (dictionary), the optional version of the scheme which one is show below:

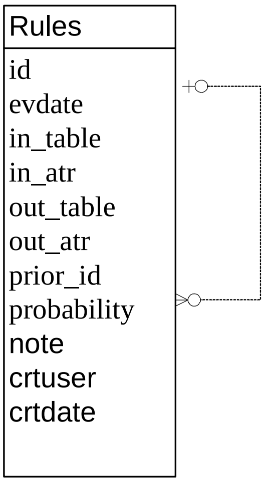

Fig. 1. The scheme of Rules relation

The scheme of relation: ID - code, EVDATE - date of a urgency of a rule, IN _ TABLE - title of the table - ancestor, IN _ ATR - attribute - ancestor, OUT _ ATR attribute - child, OUT _ TABLE - table - child, PRIOR _ ID - foreign key of table Rules (for formation of the rules with the compounded(drawn up) parts of the ancestors or child's), PROBABILITY - confidence to the rule.

The modified sweep method in turn sorts out all rules from relation Rules and applies it to tuples of relations indicated in the conforming tuple by the rule, which one is applied.

5. Conclusions

The processing of uncertainty is the key moment for many recovery methods of the data. The existing methods of elimination of equivocations process only absence, incompleteness and illegibility.

Scientific novelty. In the article the model of the class and classification rules as indistinct functional connections is offered. The methods of definition of a measure of confidence to objects of classes are tendered.

Practical value. The scientific outcomes obtained in the given article, resolve to conduct further practical researches on discriminatory analysis with the purpose of elimination of uncertainty. It is offered to determine classification rules by the analysis of the existing rules.

References

[1] Panti, G. Multi-valued logics, in: D. Gabbay, P. Smets (eds.) Handbook of Defensible Reasoning and Uncertainty Management Systems. vol. 1: P. Smets (ed.) Quantified Representation of Uncertainty and Imprecision. Kluwer Acad. Publ., Dordrecht. – 1998. – P. 25-74

[2] Д.Мейер Теория реляционных баз данных: Пер. с англ.- М.: Мир, 1987. – 608 с., ил.

[3] Huhtala Y., Karkainen J. Tane: An Efficient Algoritm for discovering Functional and Approximate Dependencies// The Computer Journal. 1999. – Vol. 42. - № 2.

[4] Шаховська Застосування багатозначної логіки у базах даних. Вісник НУ ЛП № №386, 2000.