THE MATHEMATICS BEHIND *WIZRULE*

This appendix briefly presents a mathematical formulation of the rules and the deviations. It is addressed to those who are interested in mathematics. If you are *not* interested in mathematics, simply skip this appendix. You don't have to be a mathematician in order to use *WizRule* effectively.

WHAT TYPES OF RULES *WIZRULE* REVEALS?

This section provides a formal definition of the rules that *WizRule* reveals. Each rule contains a dependent variable. In if-then rules, this is the field in the "then" part. In formula rules, this is the field (variable) on the left side of the formula. The dependent field (variable) is denoted here by *R*.

IF-THEN RULES

WizRule searches for if-then rules having the following structure:

> If (condition)
> Then *R is r*
>
> Rule's probability is *p*
> The rule exists in *m* records.
> Significance level: Error probability is α

A condition in the "If" clause can be single or composite. The single condition is of the following type: the value of field *X* is *a* (if the field *X* is qualitative) or the value of field *X* belongs to interval $[x_1, x_2]$ (if the field *X* is quantitative).

A composite condition is a conjunction of several single conditions. The number of different single conditions included in a composite condition is restricted only by the number of fields in the file. An example of a composite condition containing three single conditions is:

The value of qualitative field X is a,
and the value of qualitative field Y is b,
and the value of quantitative field Z belongs to $[z_1, z_2]$.

The rule's conclusion (the "then" part) has the following type: R *is* r (if field R is qualitative), or $R \in [r_1, r_2]$ (if R is quantitative).

The quantitative variables (fields) are segmented into intervals. When constructing these intervals *WizRule* analyzes the corresponding frequency distribution function and determines the intervals with a great density of the variable's values. The number of cases belonging to each interval must be no less than the given minimum number of cases in a rule.

IF-AND-ONLY-IF RULES

When there is a one-to-one correspondence between the values of two fields, X_1 and X_2, the structure of the rule is:

The value of X_1 is a
if and only if the value of X_2 is b.

Unlike other rules, the rule probability in this case is always 1.

RULES REFERRING TO CHARACTERS AT THE BEGINNING OF A FIELD

WizRule can analyze not only the values of a field as a whole but the internal characters in these values as well. The purpose here is to find a correlation, such as that between the first digits of a phone number and an address. This is achieved by searching for rules with the following structure:

If the value of X_1 starts with the string $c_1\ c_2\ c_3 \ldots$
then the value of X_2 is b.
and
If the value of X_1 is a,

then the value of X_2 starts with the string $c_1\ c_2\ c_3\ \ldots$

where the string $c_1\ c_2\ c_3\ \ldots$ may contain from 1 to 5 characters.

RULES REFERRING TO DATE FIELDS

Date fields are considered either quantitative or qualitative. All the rules described previously apply to date fields as well. However, since the date can be analyzed further, *WizRule* searches for the following additional rules:

Rules referring to the *day of the month*:

- If the value of date field X_1 contains the *n*th day of the month, then the value of X_2 is *a*
- If the value of X_1 is *a,*
  then the value of date field X_2 contains the *n*th day of the month.

Rules referring to the *day of the week*:

- If the value of date field X_1 contains the *n*th day of the week, then the value of X_2 is *a*.
- If the value of X_1 is *a*
  then the value of date field X_1 contains the *n*th day of the week.

Rules referring to the *month of the year*:

- If the value of date field X_1 contains the nth month of the year,
  then the value of X_2 is a.

- If the value of X_1 is a,
  then the value of date field X_2 contains the nth month of the year.

Rules referring to the *year*:

- If the value of date field X_1 contains the year n,
  then the value of X_2 is a.

- If the value of X_1 is a,
  then the value of date field X_2 contains the year n.

Unconditional Rules

WizRule also searches for unconditional rules of the following structure:

The value of field X is a_1 or ... or a_k .
The rule exists in m records.

The frequency f_i of value a_i must satisfy the condition:

$$f_i \geq \max(0.1\,N, N_0)$$

where N is the number of records in the data and N_0 is the given minimum number of cases in a rule. $k \leq 10$.

The necessary and sufficient condition for establishing a rule of this type is:

$$\sum_{i=1}^{k} f_i \geq 0.99 \cdot N$$

Formula Rules

For numeric fields, *WizRule* searches for formulas in which one of the fields is a dependent variable. In general, formula rules are the following statements:

Field R is the certain function of the other fixed field(s).

Accuracy level is ρ

The rule exists in *m* records.

The following is a list of formulas that *WizRule* searches. In this list, X_1, X_2, X_3 and so on denote numeric fields (variables), while *a* and *b* denote fixed values (constants). *WizRule* does not look for all possible types of formulas; but rather restricts the search to those formula rules that may be expected in data sets of ERP systems.

ARITHMETICAL RELATIONSHIPS BETWEEN FIELDS

WizRule searches for all formulas (functions) of the following type:

$$R = (X_1 \oplus X_2) \otimes (X_3 \oplus X_4)$$

where:

$\oplus$ denotes addition (plus) or multiplication

$\otimes$ denotes one of the following four operations: addition (plus), subtraction (minus), multiplication and division

Notes:

The operation in ($X_1 \oplus X_2$) may be identical to or different from the operation in ($X_3 \oplus X_4$).

The variables X_2 and/or X_4 may be absent from the formula. Thus, the above formula may be a function of two, three or four variables.

A field may reappear twice in the formula. In such a case, *WizRule* simplifies the formula, for example: R = (X_1 + X_2) /2.

RULES REFERRING TO PERCENTAGE

In the previous formula, ($X_3 \oplus X_4$) may be substituted by one of the following two functions:

$$(1 - X_3 / 100)$$

$$(1 + X_3 / 100)$$

As a result, *WizRule* can reveal a formula such as:

$$R = X_1 \cdot X_2 \cdot (1 + X_3 / 100)$$

This formula is efficient when the value of X_3 is a percentage. The type (unit measure) of X_3 should be Number.

LINEAR FUNCTIONS

$$R = a \cdot X_1 + b$$

HYPERBOLIC FUNCTIONS

$$R = a / X_1$$

Spelling Rules

WizRule searches for spelling rules for each qualitative field separately. Spelling rules are the following statements:

> The value $c_1\ c_2\ c_3$ appears N times in qualitative field X.
>
> There are M cases containing similar value(s) having a very low frequency (i.e., not greater than 3).

Such a rule must satisfy the following conditions:

> The length of value $c_1\ c_2\ \ldots\ c_n$ $n \geq 4$.
>
> $N \geq \max(N_0, 20)$, where N_0 is the given minimum number of cases in a rule.
>
> $M \leq \frac{1}{3} N$, where M is the sum of frequencies of infrequent similar values. The notion "infrequent value" has been defined previously.

The definition of a "similar value" is as follows:

The value $a_1\ a_2\ \ldots\ a_m$ is similar to value $c_1\ c_2\ \ldots\ c_n$ if and only if one of the following four conditions is fulfilled:

$m = n$, and $\exists k$

$\mid \quad a_k = c_{k+1}\ ; \quad a_{k+1} = c_k\ ; \quad a_i = c_i\ , i \in \{1, \ldots, n\} \setminus \{k, k+1\}$.

$m = n$, and

$\exists k \quad \mid \quad a_i = c_i\ , \quad i \in \{1, \ldots, n\} \setminus \{k\}\ ; \quad a_k \neq c_k$.

$m = n - 1$, and $\exists k \quad \mid \quad a_i = c_i\ , \quad i = 1, \ldots, k-1\ ; \quad a_i = c_{i+1}\ , i = k, \ldots, m$.

$m = n + 1$, and $\exists k \quad \mid \quad a_i = c_i\ , \ i = 1, \ldots, k-1\ ; \quad a_{i+1} = c_i\ , i = k, \ldots, n$.

THE RULE ERROR PROBABILITY

In addition to the "if" and "then" clauses, each if-then rule has the following parameters:

- The rule probability
- The rule exists in *m* records
- The significance level: error probability of the rule

The expression "The rule exists in *m* records" means that *m* records satisfy both the rule's condition *and* the rule's conclusion. Some data mining literature uses the term "support level" rather than "number of cases in a rule." However, the two terms are synonymous.

The *rule's probability p* is the probability that for a random record satisfying the rule's condition, the rule's conclusion is also fulfilled. Therefore, $p = m / n$, where *n* is the number of records satisfying the rule's condition. Some data mining literature uses the term "confidence level" rather than "rule probability." However, the two terms are synonymous.

The *significance level* of the rule designates the probability that the rule does not exist accidentally. It can be interpreted as a measure of the rule's validity. The significance level is equal to $1-\alpha$, where α is the *error probability*.

To precisely define error probability, the following notations will be used:

m is the number of cases in a rule
p is the rule's probability
n = m / p is the number of records satisfying the rule's condition
α is the a priori probability (error probability) of the rule

Consider an if-then rule, where *R is r* is in the "then" part. Let *q* be the frequency of the value *r* (in the field *R*) within the data, and N the number of records in the data. Therefore, $a = q / N$ is the average probability that the field *R* includes the value *r*.

Consider the rule's probability as a random variable, denoted by *x*. We will assume that random variable *x* is normally distributed (according to Gauss distribution) with the statistical mean *a*. The standard deviation σ of this normal distribution is chosen so that the a priori probability α for the appropriate rule with $p = 1$ is equal to $\frac{1}{n+1}$, if $a \geq 0.5$. If $a < 0.5$, then σ is chosen so that α for the rule with $p = 0$ is equal to $\frac{1}{n+1}$.

Let $\Phi_0(x)$ be the distribution function for the normalized and centralized normal distribution; $\Phi_0(x) = \frac{1}{2\pi} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$.

Following the definition of the a priori probability of a rule, α is calculated as follows:

Find σ:

$$\sigma = \Phi_0^{-1}(0.5 - \frac{1}{n+1})$$

Calculate *b*:

$$b = \frac{\sigma \cdot |p-a|}{1-a} \text{ , if } a \geq 0.5;$$

$$b = \frac{\sigma \cdot |p-a|}{a} \text{ , if } a < 0.5.$$

Determine Φ_0 (*b*).

α = 0.5 - Φ_0 (*b*)..

ACCURACY LEVEL OF FORMULA RULES

For any numeric field *R*, *WizRule* searches for the functions between field *R* and other numeric fields. The types of functions that *WizRule* searches for are defined previously. For definiteness, consider the function of three variables *R = f (x, y, z)*. The accuracy level ρ for function *R = f (x, y, z)* is calculated as follows.

The value of field *R* at record *i* is denoted by r_i, *i = i, …, N*, where *N* is the number of records in the data. Let x_i, y_i, z_i be the values of the variables *x, y, z* respectively at record *i* .Then

$$\rho = \frac{n}{N} ,$$

where *n* is the quantity of records at which the equality

$$r_i = f\ (x_i\ , y_i\ , z_i\ )$$

is exactly fulfilled.

The formula is accepted as a rule if $p \geq p_{\min}$, where $p_{\min}$ is the given minimum accuracy level.

According to the above definition, any deviation (independently, whether great or small) has the same influence on the accuracy level of the formula. Thus, any deviation from the formula is a case to be audited.

THE DEVIATION'S LEVEL OF UNLIKELIHOOD

The level of unlikelihood for each deviation is determined after all the rules have been revealed. It is calculated as follows:

THE LEVEL OF UNLIKELIHOOD OF A DEVIATION FROM ONE IF-THEN RULE

Consider the following if-then rule:

> If (condition)
> then *R* is *r*
>
> *Rule's probability is p*
> *The rule exists in m records.*
> *Significance level: Error probability is* α

Let r_1 be a value of the field *R* in a record deviating from this rule. We will use the following notations:

> *x* is the frequency of value r_1 of field *R* under the rule's condition;
> *y* is the frequency of this value within the entire investigated file;

n is the quantity of records satisfying the rule's condition;

$$n = \frac{m}{p}$$

N is the number of the investigated file's records.

The record-deviation is considered a case to be audited only if the following inequality is fulfilled:

$$\frac{x}{y} < \frac{n}{N}$$

For measuring the deviation $\frac{x}{y}$ from $\frac{n}{N}$, the following coefficient ρ is applied:

$$\rho = \frac{x \cdot N - n \cdot y}{x \cdot (N + x - n - y) + (n - x)(y - x)}$$

The record-deviation is a case to be audited only if $p < 0$.

We will assume that the level of unlikelihood for the considered record-deviation is equal to the significance level of the rule (that is, 1 - α, where α is the error probability of the rule) if $p = p_0$, where

$$\rho_0 = \frac{M \cdot n - m \cdot N}{m \cdot (N + m - n - M) + (M - m)(n - m)}$$

Here *M* is the frequency of value *r* of the field *R* within the data file. It is obvious that $p_0 < 0$. This follows from the fact that for any rule the inequality $\frac{m}{n} > \frac{M}{N}$ must be fulfilled.

If the coefficient ρ is negative but extremely near 0, then it is natural to suppose that the corresponding level of unlikelihood is equal to 0.5. Thus,

the level of unlikelihood is the decreasing function of ρ. This function is defined on the half-open interval [-1, 0) and has values within the interval (0.5, 1].

THE LEVEL OF UNLIKELIHOOD OF A DEVIATION FROM ONE FORMULA RULE

Since each of the formula variables (fields) includes at least five different values, the level of unlikelihood *P* for a deviation from a formula is calculated as follows:

$$P = \frac{(1-\rho)\cdot z + N\cdot\rho - 5}{N-5},$$

where:

z is the number of different values in dependent field *R*;
ρ is the accuracy level of the formula rule;
N is the number of records in the investigated file.

THE LEVEL OF UNLIKELIHOOD OF A DEVIATION FROM SEVERAL RULES

Let P_k be the level of unlikelihood for this record calculated for the rule *k*. The total level of unlikelihood *P* is calculated as follows:

$$P = 1 - \prod_k (1 - P_k)$$

Note that if a record is a positive example of at least one rule, it is *not* considered as a case to be audited even if it deviates from other rules.