Rough Dependencies as a Particular Case of Correlation: Application to the Calculation of Approximative Reducts*

María C. Fernandez-Baizán[1], Ernestina Menasalvas Ruiz[1]
José M. Peña Sánchez[1] Socorro Millán[2], Eloina Mesa[2]

[1] Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software, Facultad de Informática,U.P.M., Campus de Montegancedo, Madrid
[2] Universidad del Valle, Cali. Colombia
{cfbaizan, emenasalvas}@fi.upm.es, chema@pegaso.ls.fi.upm.es, millan@borabora.edu.co, eloimesa@pino.univalle.edu.co

Abstract. Rough Sets Theory provides a sound basis for the extraction of qualitative knowledge (dependencies) from very large relational databases. Dependencies may be expressed by means of formulas (implications) in the following way:

$$\{x_1, \ldots, x_n\} \Rightarrow_\rho \{y\}$$

where $\{x_1, \ldots, x_n\}$ are attributes that induce partitions into equivalence classes on the underlying population.
Coefficient ρ is the dependency degree, it establishes the percentage of objects that can be correctly assigned to classes of y, taking into account the classification induced by $\{x_1, \ldots, x_n\}$. Dealing with decision tables, it is important to determine ρ and to eliminate from $\{x_1, \ldots, x_n\}$ redundant attributes, to obtain minimal reducts having the same classification power as the original set. The problem of reduct extraction is NP-hard. Thus, approximative reducts are often determined. Reducts have the same classification power of the original set of attributes but quite often contain redundant attributes.
The main idea developed in this paper is that attributes considered as random variables related by means of a dependency, are also correlated (the opposite, in general, is not true). From this fact we try to find, making use of well stated and widely used statistical methods, only the most significant variables, that is to say, the variables that contribute the most (in a quantitative sense) to determine y.
The set of attributes (in general a subset of $\{x_1, x_2, \ldots, x_n\}$) obtained by means of well-founded sound statistical methods could be considered as a good approximation of a reduct.

Keywords: Rough Sets, Rough Dependencies, Multivariate Analysis, Multiple Regression.

* This work is supported by the Spanish Ministry of Education under project PB95-0301

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 335–340, 1999.

1 Rough Dependencies Reducts

Let $U = \{1, 2, \ldots, n\}$ be a non empty set of objects that will be called the *universe*. Objects of the universe are described by means of a set of attributes: $T = \{x_1, x_2, \ldots, x_k\}$.
If we assume all these attributes to be mono valued functions of the elements of U, then they can be seen as equivalence relations on U. The corresponding quotient sets being:

$$U/x_j = \{[i]_{x_j}/i \in U\} \tag{1}$$

where $[i]_{x_j}$ stands for the equivalence class (with respect to x_j) including the element i,.
Let $P \subset T$ be a subset of T. The indiscernability relation with respect to P, $IND(P)$, is defined as follows:

$$U/IND(P) = \bigcap_{x_j \in P} [i]_{x_j} \tag{2}$$

The indiscernability relation is an equivalence relation. Let now consider the following sets: $P \subseteq T$ and $Q \subseteq T$. We say that Q depends on P, $P \Rightarrow Q$, if and only if $IND(P) \subseteq IND(Q)$ (every class of $IND(P)$ is included in a class of $IND(Q)$).
In general and due both to the random nature of data and the inherent imprecision of the measures, from a table of observations we cannot infer exact dependencies. All that can be obtained are expressions of the form: $P \Rightarrow_\rho Q$. Being ρ the dependency degree , $0 \leq \rho \leq 1$, where 1 corresponds to the total dependency and 0 to the total independency of Q with respect to P.

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{IND(P)}X \tag{3}$$

$$\underline{IND(P)}X = \cup\{Y \in U/INDP/Y \subseteq X\} \tag{4}$$

We can now define ρ as $\frac{cardPOS_P(Q)}{cardU} \times 100$, the meaning of the dependency $P \Rightarrow_\rho Q$ is that the $\rho\%$ of the elements of U can be correctly assigned to classes of Q, given the classification P.
If deleting $x_j \in P$, the equality $POS_{P-\{x_j\}}(Q) = POS_P(Q)$ holds, then we say that x_j is Q-redundant in P and it may be suppressed while preserving the classification power of he set.
If $P' \subset P$ is such that $POS_P(Q) = POS_{P'}(Q)$ and P' does not contain Q-redundant elements, then we say that P' is a Q-reduct of P. Dealing with decision tables, and being $C = \{x_1, x_2, \ldots, x_k\}$ (condition attributes) and $D = \{y\}$ (decision attribute), the dependency $C \Rightarrow_\rho D$ holds, and we must: (a) determine ρ and (b) minimise C (eliminating redundancies by means of extracting reducts from it).

2 Techniques for the Multidimensional Analysis of Data: Correlation and Multiple Regression

The choice of a statistical technique for the multidimensional analysis of data depends on the nature of them as well as on the desired objective: description or prediction.

Dealing with decision tables the problem can be seen as the prediction of the decision attribute making use of the condition attributes. We can distinguish two different cases to which regression technique is applicable:

- When the predictive variables (in our case condition attributes) are quantitatives ones and the predicted variable (the decision attribute in our case) is also quantitative.
- When the predictive variables are quantitatives and the predicted variable is qualitative but can be expressed by means of a numerical value with a logical order.

3 Multiple Regression

In simple correlation there is only one predictive variable and one predicted variable. The n available examples constitute a cloud of dots in the two dimensions plane (X, Y) through which the minimal square straight line is drawn. In multiple regression this procedure is generalised. Having k predictive variables we have to calculate k coefficients $A_1, A_2, \ldots, A_k$ as well as a constant term y_0 that allow you to form the equation:

$$y = y_o + A_1x_1 + A_2x_2 + \ldots, +A_kx_k \tag{5}$$

of the regression hyperplane that approximate the best the n examples.
Assuming n to be $n >> k$: The k coefficients determine a vector A and the values of $x_1, \ldots, x_k$ constitute a matrix $X(n, k)$. The n values of Y form a column.

$$A = \begin{pmatrix} A_1 \\ \cdot \\ \cdot \\ A_k \end{pmatrix} \qquad Y = \begin{pmatrix} Y_1 \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} \tag{6}$$

$$\text{and we get :} \quad A = (X'X)^{-1}X'Y \quad (\text{Being } X' \text{ the transpose of } X) \tag{7}$$

In order to calculate these coefficients the method of centred variables is applied. To evaluate the quality of the approximation the difference between the observed and the predicted values is calculated. Let s be the result of adding the square of that difference. We define then:

$$\sigma^2 = \frac{s}{n - k - 1} \tag{8}$$

When $n >> k$ this value is approximately the variance of sample of the n examples. Then we have the correlation coefficient r to be:

$$r = \sqrt{1 - \frac{s}{\sum (y_i - \overline{y})^2}} \tag{9}$$

being $-1 \leq r \leq 1$. We consider values $|r| > 0,8$.

4 Stepwise Regression

We are interested only in the most significant variables "explaining" or "predicting" Y. To eliminate the less significant ones, we follow an iterative process of stepwise regression.
The steps are the following:

- Carry out the simple regression process with every variable under consideration. Then, retain the one giving the maximal value of r (or the minimal value of s).
- Carry out double regression process with the selected variable and any other one. Retain the one giving minimal value of s.
- We follow in this way, (triple regression, ...) In each step there is a decrement δ of s. We calculate:

$$F = \frac{\delta}{\sigma^2} \tag{10}$$

  We compare this result with the value given by a Fischer table for $(n-k-1)$ and 1 degree of freedom. We finish when the result of this test is negative $(F_{calculated} < F_{Givenbythetable})$.

The set of condition variables selected in this way is an approximative reduct.

5 Correlation vs Rough Dependencies

Correlation does not mean causality. If two independent variables depend on a third one, they will be strongly correlated. Correlation does not implies dependency. But if y depends on $\{x_1, x_2, ..., x_k\}$; then $\{y\}$ will be correlated with $x_1, x_2, ..., x_k$.

6 Stepwise Regression as a Foundation for the Calculation of Approximative Reducts

When dependencies as $\{x_1, x_2, ..., x_k\} \Rightarrow \{y\}$ are simplified, we consider possible dependencies existing between subsets of $\{x_1, x_2, ..., x_k\}$ and thus eliminating redundancy.

The statistical approach is similar: If there is a set of variables strongly correlated in the implicant, there is redundant use of the less significants, that may be eliminated by means of the stepwise regression. The subset of condition variables obtained in this way is an approximative reduct.

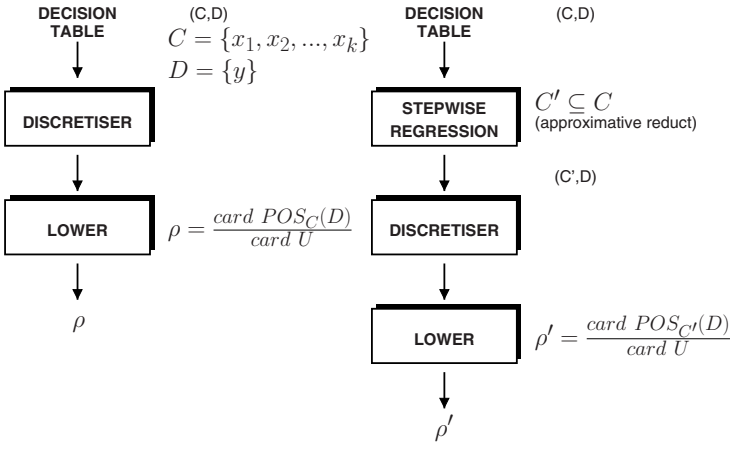

Fig. 1. Approximative reducts by means of Stepwise regression.

7 Calculating Approximative Reducts by Means of Stepwise Linear Regression

In a decision table when: (i) The number of cases (rows) is much more greater than the number of attributes (columns), (ii) The condition attributes are quantitative and (iii) There is either a qualitative or quantitative susceptible of being expressed as a numerical value with an order decision attribute. The dependency between condition and decision may be analytically approached by means of a linear regression model[1]:

$$y = y_0 + A_1 x_1 + ... + A_k x_k \tag{11}$$

The stepwise regression process provides for the elimination of the less significant condition attributes, thus obtaining an approximative reduct, whose *quality* may be tested by comparing the percentage of objects classified by using the whole set of conditions.

8 Application to Randomly Generated Data

A table containing 10.000 tuples has been generated in a random way. The table corresponds to a decision table composed by 4 condition attributes x_1, x_2, x_3, x_4 and one decision attribute y. The correlation matrix is:

$$\begin{pmatrix} 1 & & & & \\ 0,731 & 1 & & & \\ 0,816 & 0,229 & 1 & & \\ -0,535 & -0,824 & -0,139 & 1 & \\ -0,821 & -0,245 & -0,973 & -0,029 & 1 \end{pmatrix} \tag{12}$$

[1] If $|r| < 0.8$ the linear model is not adequate and other approaches should be used.

The selection of variables (in the order x_4, x_1, x_2, x_3) results:

$$y = 117,57 - 0,738x_4 \qquad (r = -0,821; \delta = 0,674) \tag{13}$$

$$y = 103,1 - 0,614x_4 + 1,44x_1 \qquad (r = 0,986; \delta = 0,297) \tag{14}$$

$$y = 71,65 - 0,237x_4 + 1,452x_1 + 0,416x_2 \quad (r = 0,991; \delta : \text{non-sensitive}) \tag{15}$$

The correlation matrix indicates that there is a high correlation index between x_2 and x_4. The minimum distance between the correspondent coefficient and its standard deviation, pointed out the need to eliminate x_4. Thus, the following result is obtained as a lineal model of y:

$$y = 52,58 + 1,468x_1 + 0,662x_2 \quad r = 0,989 \tag{16}$$

Considering the possible existence of a dependency between $\{x_1, x_2, x_3, x_4\}$ and $\{y\}$ you get that an approximative reduct is $\{x_1, x_2\}$. This method has the advantage, from the point of view of minimising the error, of calculating approximative reducts from raw (non discrete) data. However, if we apply a discretising method and then calculate the dependency degree making use of rough sets we obtain the following result:

$$\{x_1, x_2, x_3, x_4\} \Longrightarrow_{\rho_1} \{y\} \quad \rho_1 = 88,27\% \tag{17}$$

$$\{x_1, x_2\} \Longrightarrow_{\rho_2} \{y\} \quad \rho_2 = 86,74\% \tag{18}$$

From this result we can conclude that the power of classification remains almost unalterable.

References

1. Cristine Nora *Analyse de Donnees et Information*, Ecole Nationale Superieure des Telecomunications. Paris. Research Report ENST C - 79022
2. Sergey Brin, Rajeev Motwani,Craig Silverstein *Beyond Market Basket: Generalizing Association Rules to Correlations*, In Proceedings of ACM SIGMOD International Conference 1997 pp. 265-276
3. M. Jambu *Classification automatique pour l'analyse des donnes* , Vol.1 Methods et Algorithms ed. Dunod Paris 1978
4. I. C. Lerman *Classification et Analyse Ordinal des donnes. ed. Dunod 1981*,
5. Jean de Lagrade *Initiation A L'Analyse des Donnees* Dunod 1983
6. Cristine Nora, Christine Vercken *Panorama Des Principales Techniques D'Analyse De Donnes Multidimensionnelles et De Leurs Possibilites* Ecole Nationale Superieure des Telecomunications. Paris. Research Report ENST C - 76010
7. Pawlak *Rough Sets: Theoretical Aspects of Reasoning about Data* Kluwer 1991
8. Grizzle, J.E., Williams, O.D. [1972] *Loglinear models and test of independence for contingency tables. Biometrics 28, pp.137-156*
9. Bishop, Y., Fienberg, S., Holland P. *Discrete Multivariate Analysis:Theory and Practice.* Cambridge, MA: The MIT press,Second Printing , 1975
10. Agresti, A. *Analysis of ordinal caetgorical Data. Jhon Wiley and Sons, Inc. New York 1984*
11. Cox, D.R. *The Analysis of Binary Data, New York, Halsted Press, 1970*