A LOGIC-BASED EXPERT SYSTEM FOR MODEL-BUILDING IN REGRESSION ANALYSIS

Ferenc Darvas, Kornél Bein, Zoltán Gabányi
Company for Computer-Assisted Drug Design*,
1054 Budapest, Akadémia u. 17.

1. Introduction

Methods of mathematical statistics and pattern recognition make a significant part of computer applications besides data processing. Although there are numerous expert systems based on these methods /REG81/, relatively few attempts have been made for automatic building of models, computations to be based on, as well as for automatic evaluation of results, which form a major part of brain-work.

This situation seems to be apt also to regression analysis, the most wide-spread method of mathematical statistics which has been referred to in altogether two publications on possible automation.

In this paper application of logic programming in automatic model-building for regression analysis is presented, in connection with a drug design problem. After a survey of the regression problem /Chapter 2/, the drug design problem and the logical model for problem solving will be dealt with /Chapter 3/. Chapter 4 gives a summary of the program system implementing the model, while Chapter 5 reports on experiences with the system and evaluates the chosen logic programming method.

---

* A joint company of the Institute for Coordination of Computer Techniques, Budapest, and the Institute of Enzymology, Biol. Res. Cent., Hungarian Academy of Sciences, Budapest.

2. The regression problem

It is a frequent case that in regression equations of the form

$$\bar{y} = \bar{\bar{x}}\,\bar{b} + \bar{\varepsilon} \qquad /1/$$

some of the $\bar{x}_i$ column vectors of $\bar{\bar{x}}$ matrix are binary vectors, i.e. the components of x_i can take the discrete values of 0 or 1 only /DRA66/. Components of such vectors can be regarded as logical variables of values "true" or "false" and can be subjected to Boole /AND, OR, NOT/ operations.

Suppose that a variable x_i is assigned to the column vector $\bar{x}_i$, and values of 1 and 0 of x_i correspond to the presence and absence of β_i.

If the model permits to interpret physically a common variable for β_i and β_j, a new vector $\bar{\eta}$ can be introduced, as the logical OR of x_i and x_j:

$$\bar{\eta} = \bar{x}_i \quad \text{OR} \quad \bar{x}_j = \bar{x}_i + \bar{x}_j \qquad /2/$$

More generally

$$\bar{\eta} = \sum_{k=1}^{l} \bar{x}_k \qquad /3/$$

Similarly, a new vector can be generated by performing the Boolean AND operation on original vectors x_k:

$$\bar{\tau} = \prod_{k=1}^{l} \bar{x}_k \qquad /4/$$

Logical combinations according to /3/ and /4/ might be useful in all cases when x_i-s are not competely independent of each other /as it is the common case in many fields/. Causal interpretation of the regression equation requires, however, that each variable,

formed by Boolean operation, should have a meaningful interpretation in the frame of the model investigated.

3. Prediction of drug activity

Drug design aims at predicting biological activity of not yet synthetized or, at least, not yet tested compounds /HAN69/. In its most widely applied approaches /HAN63, FRE 64/, linear relationships between two groups of quantitative descriptors are searched. The first group relates to the biological activity, the second one to the chemical structure of a series of organic compounds. An important group of chemical descriptors is formed by the so-called "indicator" variables, giving the presence or absence of definite groups within the molecules /MAR78/. In drug design methods like in the Free-Wilson approach, Fujita-Ban approach and Kubinyi's mixed method, eq. 1. comprises exclusively or additionally indicator variables as $\bar{x}_i$ column vectors of $\bar{\bar{x}}$. /FRE64, FUJ71, KUB763/.

Indicator variables in eq. 1. can be interpreted as logical variables and also combined in sense of eq. /3/ or /4/ /GOL80/. Because of the high number of the possible logical combinations, the regression eq. 1. cannot be solved with a pre-determined set of all combined variables. Input variable set of the normally used stepwise regression program might include only those logical combinations, which can be interpreted in the context of the biological activity investigated.

Interpretation of the large quantity /sometimes several hundred/ of combined variables means a formidable work, which is burdened with errors of subjective decisions. We think that the high intellectual expenditures of such interpretations compose the main reason for the fact that logical combinations are rarely used in drug design calculations /HAN75, ELG82/, though the first

examples were published a long time ago /BOC65, KOP65, SCHA75, KUB762/.

In order to find physical interpretation for the combined variables by logical programming, a logical model of the drug-receptor "reaction", the ultimate scene of the drug action is needed. Here we give only an informal summary of the most important concepts we have used in our model.

It is supposed that all compounds act with the same "reaction mechanism" on the same macromolecule /receptor/ within a living cell. Measured biological activity values used in eq. 1. are originated almost exclusively from this reaction.

Structure of the compound series can be described as aggregate of unique groups /occuring only in some compounds/ and of the remaining part of the molecule /supposed to be common in all compounds/. Indicator variables of eq. 1. denote the presence or absence of single groups in each compound.

Contributions of the i-th and the j-th indicator variables / γ_i and γ_j/ to the biological activity of all compounds are expressed as the regression coefficients b_i and b_j. b_i and b_j depend, in the first place, on the sets of chemical and physical properties /ε_i and ε_j/ of the chemical groups β_i, and β_j.

There are two cases:
1. Contributions of γ_i and γ_j are independent from each other.
2. Contributions of γ_i and γ_j depend on each other.

In this latter case, it is supposed that γ_i and γ_j enter into an "interaction". Such interaction can be originated e. g. from the formation of an intramolecular chemical bond between the

substituents /FUJ71/. An interaction between γ_i and γ_j depends on the "environmental" conditions in addition to the properties of ε_i and ε_j. Most important groups of the environmental conditions are the

- electronic connections between two groups, transmitted through the common part of all molecules, and
- through-space connections between two groups being able to reach each other.

Let describe δ_{ij} and μ_{ij} the two sets of connections between γ_i and γ_j.

The sets $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_Q$ are constructed in such manner, that they give a possible full description of the meanings of $\gamma_1, \gamma_2 \ldots \gamma_Q$ groups, attaching all to the same position (P_1) of the common part of molecules.

A variable resulted from logical addition of $\gamma_1, \gamma_2, \ldots \gamma_N$, where $N<Q$ according to eq. 3. can be interpreted as the largest common subset within $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_N$ which, in the same time, does not occur in the $\varepsilon_{N+1}, \varepsilon_{N+2} \ldots \varepsilon_Q$ property sets of the remaining $\gamma_{N+1}, \gamma_{N+2}, \ldots \gamma_Q$ groups attaching to P_1.

Logical multiplication of γ_1 and γ_2 according to eq. 4. can be interpreted as an indication to an interaction, I_1.
I_1 occurs if properties of γ_i and γ_j, $\varepsilon_i' \in \varepsilon_i$ and $\varepsilon_j' \in \varepsilon_j$ enable chemical or physical "modification" of β_i and β_j trough the connections $\mu_{ij}' \in \mu_{ij}$ and $\delta_{ij}' \in \delta_{ij}$.

Thus, they can be represented in form of a production rule /SHO76/

$$\varepsilon_i', \varepsilon_j', \delta_{ij}', \mu_{ij}' \longrightarrow I_1 \qquad /5/$$

Subsets of properties can cause several interactions. On the other hand, an interaction can be triggered by several combination of property sets.

4. Computer implementation of the model

Biological activities and chemical structural descriptors of the tested compound are the starting data of the calculations. Structural descriptors can be divided into two classes: binary indicator variables and continuous physico-chemical, quantum-chemical variables. Values of these continuous variables can be calculated by other programs or retrieved from the data base stored in the system. They can be transformed in different ways /addition, subtraction, taking logarithm/. Structural formulae of the compounds with activities to be predicted also belong to the starting data set.
First step of the program is to solve a so-called Fujita-Ban equation system /FUJ71/ using the measured activities and the indicator variables. The solution serves as input for the logical interpretation of the possible logical additions.

In the PROLOG program providing the interpretation, each chemical group /"substituent"/ is characterized by a chemical property set. If a contraction /i.e. logical addition/ is carried out, common part of the property sets of the groups is generated. A contraction is permitted only if the property set of the other chemical groups at the same substitution site do not imply this common part. As interpretation of the contracted variables the resulted common property sets are considered.
After the evaluation of the possible interpretations the computation goes on in an interactive way so that the user decides on the contractions of the substituents step by step. After the user's giving two or more substituents to be contracted, the system computes the new regression equation, its statistics and optionally the estimated biological activities of the untested compounds. If the user accepts the equation, the next contraction will be carried out on the base of the new indicator variable set, otherwise the processing continues

using the previous variable set. Contractions of indicator variables are carried out at each substitution site, one by one.

Having performed all necessary contractions, generation of interaction variables follows. An interaction variable corresponds to the common presence of substituent M at site I and substituent N at site J /logical 'AND' relation/.
Value of this variable is 1 for a given compound if this 'and' relation is true, otherwise it is 0.The system generates all interaction variables with value = 1 for two or more tested compounds, and calculates frequencies for them.

The user gives a lower frequency limit f_1. This specifies that henceforth interaction variables of frequency greater than or equal to f_1 are treated only.

The interaction variables are then prescreened: the program lists the variables with statistics informing about their importance. If a variable is redundant in statistical point of view, it is ignored. The accepted interaction variables are added to the set of variables, and they are also passed to the program giving automatic interpretation.

Using the built-in automatic deductive mechanism, this program tries to prove the interaction rules stored in its data base. If the proving procedure is successful, the user is informed about the result as a possible interaction. The interpreted interaction variable can be included in the input data set for the calculation of the final regression equation. This input involves not only the original, contracted and interaction variables but the continuous ones as well. There are two means for computing the final equation: interactive or automatic stepwise regression analysis. As result of the numerical calculations one gets the regression equation corresponding to the wanted quantitative structure-activity relationship,

its statistical characteristics and the estimated activities of the tested and untested compounds.

5. Experiences. Summary

The system is implemented on the Siemens 7536 computer of the Institute for Coordination of Computer Techniques, in FORTRAN and MPROLOG. MPROLOG is a modular version of PROLOG, being developed by the Institute. Besides its comfortable program development facilities it permits modular structuring of the program.

In order to test drug design performance of the system, earlier results were recalculated and new problems were solved. Recalculating one of our earlier series of structure-activity regression equations /DAR80/ resulted significant and meaningful equations in all of the 8 cases investigated. In addition, an earlier version of the system helped in a great extent to calculate quantitative structure-activity relationships for antifungal nitroalcohols /LOP83/. In summary, the mechanical interpretation of the combined variables seems to be a helpful tool in model-building for drug design.

On the other hand, drug design has been a favourite field of logical programming for a long time. Besides a system for predicting drug interaction /DAR75, FUT76, FUT771, DAR78, FUT79/, a carcinogenity prediction system /FUT771/ and a system for calculation of physicochemical properties of organic compounds /DAR782/ have been implemented. The expert system in question has an additional feature relative to them: the general nature of the problem and the model formulated permits our program system, with minor changes, to be applied in numerous other fields as well.

Among others, potential application fields are the quality control, geology, town planning, environment protection, all of them dealing frequently with regression models including yes/no variables and their combinations. The fact, that interpretation of the combined variables, a bottleneck of the model-building, could be performed by a relatively short program shows, that logical programming is a powerful tool in constructing small expert systems.

Faltam os títulos

References

BOC65
K. Boček, J. Kopecky, M. Krivucová, D. Vlachová: Experientia 2o 667 /1965/

DAR75
F. Darvas, I. Futó, P. Szeredi, in: Proc. Conf. Comp. Cyb. Methods in Medicine and Biology, p. 413, Ed. D. Muszka, Szeged, Hungary

DAR78
F. Darvas, I. Futó, P. Szeredi: Int. J. of Biomed. Comp. 9 259 /1978/

DAR782
F. Darvas, I. Futó, P. Szeredi, in: Proc. Symp. on Chem.-Struct. - Biol. Act.: Quant. Approaches, Ed. R. Franke, Akademie Verlag, Berlin, 1978.

DAR8o
F. Darvas, J. Röhricht, Z. Budai, B. Bordás, in: Chemical Structure-Biological Activity Relationships, Eds. J. Knoll, F. Darvas, Pergamon Press, London 198o, p. 25.

ELG82
J. Elguero, A. Fruchier, Affinidad 39 548 /1982/

FRE64
S. M. Free, J. W. Wilson: J. Med. Chem. 7 395 /1964/

FUJ71
T. Fujita, G. Ban: J. Med. Chem. 14 148 /1971/

FUT76
I. Futó, P. Szeredi, F. Darvas, in: Proc. Conf. Logique et Base de Données, Toulouse, ONERA, 1977, p. 18.

FUT771
I. Futó, F. Darvas, E. Cholnoky, in: Proc. 2nd Int. Congr. of the J. Neumann Society, Budapest, 1977

FUT79
I. Futó, F. Darvas, P. Szeredi, in: Logic and Data Bases, Ed. J. Minker, Plenum Press, N. Y. 1979

GOL8o
V. E. Golender, A. B. Rozenbliet, in: Drug Design, Vol. IX., p. 299, Ed. E. J. Ariens, Academic Press, N. Y. 198o

HAN63
C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger, M. Streich: J. Amer. Chem. Soc. 85 2817 /1963/

HAN69
C. Hansch: Accounts Chem. Res. 2 232 /1969/

HAN75
C. Hansch, C. Silipo, E. E. Steller: J. Pharm. Sci. 64 1186 /1975/

KOP65
J. Kopecky, K. Bocek; D. Vlachová: Nature 2o7 981 /1965/

KUB762
H. Kubinyi: J. Med. Chem. 19 587 /1976/

KUB763
H. Kubinyi, O. Kehrhahn: J. Med. Chem. 19 1o4o /1976/

LOP83
A. Lopata, F. Darvas, K. Valkó, Gy. Mikite, E. Jakucs, A. Kis-Tamás; Pestic. Sci., accepted for publication, 1983

MAR78
Y. C. Martin; Quantitative Drug Design. A Critical Introduction. M. Dekker, N. Y. 1978

REG81
J. A. Reggia: Ann. Biomed. Eng. 9 6o5 /1981/

SCH75
L. J. Schaad, R. H. Werner, L. Dillon, L. Field, C. E. Tate: J. Med. Chem. 18 344 /1975/

SHO76
E. H. Shortlife, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, N. Cohen: Comp. Biomed. Res. 8 3o3 /1975/

DRA66
N. R. Draper, H. Smith: Applied Regression Analysis. J. Wiley, N. Y. 1966