Analyzing artificial data by the GUHA method

Kamila Bendová
Mathematical Institute ČSAV, Prague

Abstract

The data in question were created by D. Pokorný, P. Vopěnka and J. Šimek in order to test the ability of routine statistical procedures to detect a satisfactorily simple casual illness on a set of hypothetical symptoms; the result of their study was rather negative. In the present article I show that reasonable iterated use of the GUHA procedure ASSOC has lead to the same definition as originally intended and I describe the trace of the process of its discovery.

1. *Introduction*

Dan Pokorný, Petr Vopěnka and Jiří Šimek studied the possibilities of statistical data analysis especially in the case of multifactorially conditioned deceases: they invented a definition of a hypothetical illness on the basis of certain symptoms and produced artificial data according to this definition. They analysed these data using ordinary statisticalmethods and showed that none of them was able either to discover the definition or at least to be close to its discovery; moreover, different statistical methods yielded different results (see [1]).

Being engaged in processing actual empirical data by the GUHA method (see [4], [5], [6]) I found challenging and promising to subject the artificial data of Pokorný, Vopěnka and Šimek to an analysing using GUHA and try to discover the definition of the hypothetical illness in this way. The authors of [1] generously gave their data at my disposal, as well as a copy of [1] with two last pages (containing the definition of the illness) deleted. These two pages (and the correct definition of the illness by the three authors) had remained unknown to me until I presented my results on a seminar.

The data had the form of two samples: the smaller one contained information about 80 healthy and 80 ill pacients: everyone having or not having some of 50 different symptoms, symptom No. 51 represented illness as a whole. Thus it had the form of a matrix of 160 rows and 51 columns. The second sample had the same form but it contained information about 8000 pacients - 4000 ill and 4000 healthy. My goal was to find the definition of the illness on the basis of the symptoms.

2. *The GUHA method*

GUHA is a method of data analysis based on logic and statistics. It processes exactly such samples - i.e. matrices corresponding to objects and

their properties (two valued variables) - and searches for non-trivial associations and relationships among the properties or their combinations. I used the procedure ASSOC. A brief description follows:

The procedure ASSOC automatically generates combinations (elementary conjunctions) of properties and searches for the associations of these new properties. I.e. it verifies hypotheses of two types in the sample:

(1) $V \approx_\alpha W$: V is associated with W on the level α where V and W are original properties or their combinations; the association is defined by a statistic and by level of importance α;

(2) $V \Rightarrow_p W$: V p-implies W where V and W are again atomic properties or their combinations and $\Rightarrow_p$ is a p-implication, i.e. it holds if at least p percent of objects having property V (in the sample) have also property W; in particular case of p = 100 it is the logical implication. Therefore V is called the antecedent and W the succedent.

All statistics are defined on the fourfold contingency table A,B,C,D where A is the number of objects satisfying V&W, B the number of objects satisfying V&¬W, similarly C for ¬V&W and D for ¬V&¬W where ¬V is non V (negation of V).

Using the control language of ASSOC, the user can fix the following:

(succedent) - ANT and SUC
- minimal and maximal length of antecedents (succedents) - MINA, MAXA, MINS, MAXS
- the allowed form of each variable (positive, negative, both) - SIGN=POS, NEG, BOTH
- the statistic defining the association or almost-implication together with its parameters (α, p) - AL, CP
- base - the minimal frequency of V&W (If the frequency of V&W is smaller than the base, the corresponding hypothesis is not considered.) - BASE

For details see [2], [3].

The output of the procedure ASSOC is a sequence of hypotheses of the form

$$V_1 \& V_2 \& \dots V_n \; Q \; W_1 \& W_2 \& \dots W_m \qquad (+)$$

where $V_i \in$ ANT, $W_j \in$ SUC, Q is some quantifier ($\approx_\alpha$ or $\Rightarrow_p$) and the hypothesis (+) is valid in the sample. Furthermore a fourfold contingency table is printed, or some other tables (row percentual, column percentual etc.).

3. *A trace of the analysis*

In contradiction to the situation of statisticians described in the paper [1] I knew that the data are artificial and that my goal was to find the definition of the illness, not only to describe the data. I knew that every meaningful definition should be reasonably graspable and thus essentially simple and reflecting internal relationship - however, I could produce it only as a disjunction of elementary conjunctions. Moreover it seemed logical to me that it is related rather with presence of some symptoms (or simultaneous presence of some and absence of others) that with their absence only. At the artificial sample these assumption are rather "psychological" but they agree with the real situation in medicine: the symptoms must be chosen in such a way that one or several of their simple combinations define an illness, i.e. that they can be used for practical diagnosis. And already the word "symptom" suggests that it is usual to diagnose the illness from their presence in positive form (even if the property itself may correspond to the owing of something).

When analysing data I proceeded as usual, however, in distinction with the case of actual data I systematically used only the implicational quantifier $\Rightarrow_p$ because it is the only one which can yield the desired formula Φ (some combination of symptoms) equivalent to the property 51, i.e. the presence of illness. This property Φ must imply the illness in 100% of cases, i.e. in the fourfold table B = 0, and moreover C = 0, i.e. A = number of all patients.

In the following I want to describe my journey to find the resulting formula, as well as various, blind alleys and useless digressions, in order to demonstrate possibilities and procedures of the GUHA method. For clarity I shall use the usual record of individual runs.

Thus in our case the output was mostly a sequence of hypotheses of the form

$$\varepsilon_1 V_1 \ \& \ \varepsilon_2 V_2 \ \& \ \ldots \ \& \ \varepsilon_n V_n \Rightarrow_p W,$$

where $V_i \in$ ANT are symptoms, $W \in$ SUC is the illness, $P' = A/A+B \geq p$ where A, B are the frequencies from the fourfold table; if B = 0 it is a logical implication.

I used the computer IBM 370/135.

First I processed the small sample.

I. Using standard statistical software I counted the basic statistics for various symptoms and I confirmed that this is artificial sample: all symptoms have similar frequencies (about 40%).

Than I used mostly the procedure ASSOC of the GUHA method with the quantifier of p-implication.

II. INPUT: ANT = 1 to 50 (all symptoms); MINA = 1; MAXA = 2; SUC = 51 (the illness); SIGN = BOTH; BASE = 5; TIME = 10 (minutes).

In this run I tried to learn something about the data structure.

The main result consists of three symptoms 1,2,5 which themselves imply the illness in more than 75% and of many two term conjunctions some of which imply the illness in 100%. (In this time I again reflected that this is an artificial case because from my experience I know that 100%-implications exist in real data only in the case of a pure casual dependence.)

Due to above considerations about the simplicity of the definition and due to my experience I was convinced that these logical implications must be somewhat connected with the pursued definitions. I chose those symptoms which were in one of the 100% implications and I began another run.

III. INPUT: ANT = 1 to 5, 14, 15, 20, 40 to 44; MINA = 1;
MAXA = 3; SUC = 51; BASE = 5; SIGN = BOTH; CP = 0.90;
TIME = 10.

I received a plenty of two-term and three-term conjunctions which p-implied (often even logically implied) the illness. I tried to find a regularity in them but with no success. I realized that because of time and space limitations (programs of the GUHA method print as many hypotheses as specified default value being 100) in the first run I obtained only an initial segment of the sequence of all hypotheses printed - the last one was 5 & 41 (the conjunctions are generated stepwise in lexicographic order). I began the following run

IV. INPUT: ANT = 5 to 50; MINA = 1; MAXA = 2; FORM = BOTH;
SUC = 51; CP = 0.90; BASE = 5; TIME = 10.

I again received a great many hypotheses, some of them two-term formulas which logically implied the illness. But still I was no wiser. It seemed to me that my hypothesis about the role of only positive occurence of symptoms was confirmed.

In the meantime I tried to analyse the large sample:

VI. Using procedure ASSOC with
INPUT: ANT = 1 to 50; MINA = 1; MAXA = 2; SIGN = BOTH;
SUC = 51; CP = 0.90; BASE = 20; TIME = 20.
I recieved no hypothesis.

VII. I returned to the small sample. I hoped that conjunctions which imply the illness have some connection with the definition but I didn't know which. Therefore I started to interest myself in disjunctions: I realized that if

$$51 \Rightarrow \phi_1 \,\&\, \phi_2 \,\&\, \ldots\, \phi_n$$

where ϕ_i are symptoms in positive or negative form then

$$\neg\phi_1 \vee \neg\phi_2 \vee \ldots\, \neg\phi_n \Rightarrow 51.$$

Thus I began with the

INPUT: ANT = 51; MINA = 1; MAXA = 1, SIGN = NEG; SUC = 1 to 5,14,15,20,40-44; MINS = 1; MAXS = 4; P = 100; BASE = 5.

I received some four-term conjunctions but they always included both negative and positive symptoms, e.g.

$$1 \vee 5 \vee \neg 15 \vee 44 \rightarrow 51$$
$$1 \vee 20 \vee \neg 44 \vee \neg 42 \rightarrow 51$$

and I again didn't know how to use this information.

VIII. Since I believed in my choice of "suspect" symptoms I tried the large sample with

INPUT: ANT = 1 to 5,14,15,20,40 to 44; MINA = 1; MAXA = 3; SIGN = (ALL) POS; SUC = 51; CP = 0.95; TIME = 20

and to my surprise I didn't received (in 20 minutes) any hypothesis, i.e. in the given time no conjunction p-implying the illness was found. It was clear that the computation for such a large model (8000 pacients) is slow and should I want to receive something I must know much better how to determine parameters.

In spite of that it seemed clear that hypotheses verified on a small sample may be false on the large one.

IX. Now I made some pencil and paper work: I took the listing of the small sample as well as all two and three-term conjunctions implying the illness and I tried to find for every object (patient) a conjunction satisfied by this object. In other words I tried to cover the whole sample by a disjunction of conjunctions implying the illness. If I covered the whole sample I would obtain a formula equivalent to the illness - but I knew that such a complicated definition cannot be the expected one (not speaking about that it wasn't verified in the large sample).

I succeeded to cover almost the whole sample with the exception of eight objects. One of them had only symptoms 21,25,32,33,37,46,48.

Already before that I was certain that I deal with only positive forms. Now I realized for the first time that a positive occurrence of two adjacent symptoms may be important (and I could make some sense of it). I wanted to cover also the remaining eight objects by some conjunctions and I took those symptoms which occur in them positively and placed them into ANT. I also lowered the base.

X. INPUT: ANT = 10 to 14,16 to 18,22,23,34,35; MINA = 1; MAXA = 3; SUC = 51; BASE = 1; CP = 0.90; SIGN = (ALL) POS.

The output was many three-term conjunctions p-implying the illness. Again I realized that the positive occurrence adjacent symptoms can be important (also from ANT = ...22,23,34,35...).

XI. I returned to the large sample: I lowered p and reduced the set ANT:

INPUT: ANT = 1 to 5,11 to 15, 41 to 45; MINA = 1; MAXA = 4;
SIGN = (ALL) POS; SUC = 51; CP = 0.80; BASE = 20;
TIME = 20.

I obtained one hypothesis with antecedent of lenght two: 4 & 5 and many hypotheses with antecedents of length three (none of them implying the illness) but due to computation time limit not all three-term conjunctions. Thus no four-term hypotheses was produced. It was obvious that there is something not accidental here.

XII. I continue on the large sample: I lowered the number of symptoms in ANT and allowed only four-term conjunctions to be produced because I knew that smaller conjunctions didn't imply the illness.

INPUT: ANT = 1,2,4,5,41 to 44; MINA = 4; MAXA = 4; SIGN = (ALL)
POS; SUC = 51; CP = 0.95; BASE = 20; TIME = 20.

Heuréka! I received four-term conjunctions implying the illness and in which I already saw a certain system:

1&2&41&42	1&5&41&42	2&5&42&43	4&5&41&42	5&41&42&44
1&2&42&43	1&5&42&43	2&41&43&44	4&5&42&43	
1&2&43&44	1&5&43&44		4&5&43&44	

The desired definition seemed to appear: in some segments always two adjacent properties are important. From the rest I suspected that about first five numbers in every decadic segment plays the role (i.e. 1,2,..5,11,12,..15,..41,42,..45) and that always the first and the fifth symptoms have also the same importancew as other adjacent pairs of symptoms. I proceeded verified it by hand (or rather by eyes) on the small sample and I started, using standard statistical software to transform the properties in order to produce new combinations, i.e. to do four-term conjunctions which imply the illness as new properties.

As it appears it was important to select small set of antecedent properties and to stop useless computation of hypotheses with shorter length of antecedent.

XIII. INPUT: ANT = 11 to 15, 20 to 25, 30 to 35; MINA = 4;
MAXA = 4; SIGN = (ALL) POS; SUC = 51; CP = 0.95; BASE = 20;
TIME = 20.

And again I received the hypotheses which I expected.

Then I started to transform. But it was more complicated that I expected: I wanted to create many new properties in order to ensure that all new properties are right, i.e. they imply the illness. And also I made a mistake: in the first "discovery" run XII I put to ANT the symptoms 1,2,4,5 because I thought these are the right ones and then I thought that the symptom number three, which didn't occur in the result, is not important. I constantly verified the new properties by runs of procedure ASSOC when new transformed properties, e.g. 52-56, were in ANT and the illness was in SUC. With the above-mentioned errors B was zero but C did not equal zero.

The resulting transformation is as follows:

Y0 = (1&2) v (2&3) v (3&4) v (4&5) v (1&5)
Y1 = (11&12) v (12&13) v (13&14) v (14&15) v (11&15)
Y2 = (21&22) v (22&23) v (23&24) v (24&25) v (21&25)
Y3 = (31&32) v (32&34) v (33&34) v (34&35) v (31&35)
Y4 = (41&42) v (42&43) v (43&44) v (44&45) v (41&45)
X52 = (Y0&Y1) v (Y1&Y2) v (Y2&Y3) v (Y3&Y4) v (Y0&Y4).

This new property X52 implied - both on the small and the large sample - the illness and that in such a way that in the fourfold table C = 0.I.e. X52 is equivalent to the illness (on the samples).

4. *Discussion*

The original description of the illness was the following (see [1]):

Region. We suppose variables X1,X2,...X5 to be cyclically ordered, i.e. the variable X1 is the neighbour both of X2 and X5. Let m be maximal number of cyclically neighouring values "one".

If m = 0,1, then the region X is intact,
if m = 2,3, then the region X is pathologically damaged,
if m = 4,5, then the region X is letally damaged.

Person. We suppose region A,B,..E (i.e. first, second,...fifth decade) to be cyclically ordered. If any region is letally damaged, a person is assumed not to exist. Otherwise: Let n be maximal number of cyclically neighbouring pathologically damaged regions, then

if n = 0,1, then the person is healthy,
if n = 2,3,4,5,6, the person is ill.

It can be seen that the two definitions - the original and mine - are equivalent. By my opinion, this investigation has shown that the GUHA method can be especially helpful in situations when routine statistical methods fail. The reader should observe the following:

(1) The GUHA package was used in its standard form, without any changes or adaptations.

(2) Intelligent use of GUHA is necessary; and to find the proper choice of parameters may be a small invention. Blind use will hardly give useful results.

(3) The described processing is an example of iterated use of GUHA procedures.

(4) Needless to say, GUHA may fail; e.g. if the definition was slightly more complicated (such that only six-term conjunctions would lead to the solution), then needs for computer time might be not realistic.

Remark. When the result was known my friend Mrs. Anna Sochorová put these data to another faster (about 7 times) computer EC 1045 and put

INPUT: ANT = 50; MAXA = 4; SIGN = (ALL) P; SUC = 51; CP = 1.00; BASE = 40; TIME = 40;

then after more than half-an-hour she received all four terms conjunctions.

To close let me say that GUHA is indeed a very useful tool for ascertaining the structure hidden in data, but it is only a tool. However, to interpret the meaning of the message conveyed by the data is a task reserved solely for a human expert.

Many thanks to my colleagues Dan Pokorný, Petr Vopěnka and Jiří Šimek who invented the problem, created the data and gladly lent them to me. Also I want to thank my friend Stanislav Přeučil who helped me with the computing on IBM 370/135.

References

[1] D. Pokorný, J. Šimek, P. Vopěnka: On the data-analytic experiment (in this volume)
[2] P. Hájek, T. Havránek: The GUHA method, its aims and techniques, Int.Journal of Man-Machine Studies 10, 3-21
[3] P. Hájek: The new version of the GUHA procedure ASSOC; in: Proceedings of COMSTAT 1984, Physics-Verlag Wien
[4] P. Hájek, T. Havránek: Mechanizing Hypothesis Formation, Springer-Verlag, Berlin-Heidelberg-New York, 1978
[5] P. Hájek, T. Havránek, M. Chytil: Metoda GUHA - automatická tvorba hypotéz, Academia Praha 1983
[6] Special issues of Int. Journal of Man-Machine Studies 10 (1978), No.1, 15 (1981), No.3

Poznámka redakce: tento článek byl připraven do tisku v návaznosti na práci kolegů Pokorného, Vopěnky a Šimka, ale v době svého vzniku nesměl být publikován. Ze známých důvodů nebyla ani práce [1] v anglické verzi publikována.