Yugoslav Journal of Operations Research
4 (1994), Number 2, 159–165

CLAS – A FORMAL AID TO DATA ELEMENTS IDENTIFICATION

Dušan MALBAŠKI, Danilo OBRADOVIĆ

University of Novi Sad, Faculty of Technical Sciences,
Trg D. Obradovića 6, 21000 Novi Sad, Yugoslavia

Abstract: The method *CLAS* (*Continuous Logic Attribute Synthesis*) is a formalisation of the standard method named *Synthesis of Attributes* used for identification of entities and attributes in information systems design. It differs from the original method in a sense that the subjectiveness in the procedure of evaluation of parameters is eliminated by introducing formalised quantitative measures for elementary criteria. Also, the three global identification criteria which may cause a contradiction are reduced to one by the use of continuous logic functions.

Key words: Data elements, information systems, identification, synthesis of attributes.

1. INTRODUCTION

One of the major issues in systems analysis, particularly for process-oriented methods is the problem of data identification, e.g. the problem of recognizing a datum as an entity or attribute. It is well known [1] that those properties are not inherent for data, meaning that a datum can be interpreted as an entity, attribute or even relationship depending on the environment.

This paper deals with a semi-formal method for data classification named *Synthesis of Attributes* [1], which is based on the statistical model of the information system analyzed. Items of data (we will name them data elements for the sake of clearness), or rather the part of them considered in our discussion, are classified into the following categories: entities, strong entity attributes and weak entity attributes. The meaning of these terms is standard and is thoroughly explained in [1]. Statistics are introduced as an aid to classification because, as mentioned, the type of data element depends on other data elements in the system and their relation the one on hand, as well as because of some uncertainty of choice which is unfortunately true

---

Paper presented at the 2nd Balkan Conference on OR, Thessaloniki, Greece, 18–21 October, 1993

for a good deal of data elements. Data elements are related to each other through processes that can be loosely defined as programs working on data set in question.

For entities, strong an weak entity attributes a set of axioms is defined, based on practical experience, which is used for identification purposes. The set consists of three axioms, one for each type of data elements. The axioms are defined in terms of three quantitative measures:

N_1 – the number of processes where a given data element is encountered,

N_2 – the number of other data elements which are simultaneously used with the given element,

N_3 – the number of other data elements which are often simultaneously used with the given data element.

Regarding the quantities N_1, N_2, and N_3 a data element is a candidate for entity if it' s N_1 and N_2 are large and N_3 is small; it is a candidate for strong entity attribute if N_1 and N_2 are small and N_3 is large; finally it is a candidate for weak entity attribute if all values are medium.

It is readily noticed that the base of the method incorporates four measures whose values are not precise: here we have in mind the terms "often" (in the definition of N_3), as well as "large", "small" and "medium" found in axioms. The mere nature of those measures clearly suggests that the evaluation process must be in some way subjective and the aim of the original *Attribute Synthesis method* was do decrease the level of subjectiveness as much as possible. It is done through the use of histograms and visual judgement based on them. Here is a brief description of the original method, [1].

Let $(e_1, \ldots, e_k)$ and $(p_1, \ldots, p_n)$ be the sets of data elements and processes respectively. We start with a matrix M, $k \times n$, the element (i, j) of which is equal to 1 if data element e_i is used in the process p_j and 0 otherwise.

Next, we define a set of three vectors, the first being a vector v with k elements where

$$v[i] = M[i, 1] + \ldots + M[i, n], \qquad i = 1, \ldots, k.$$

Apparently, each element of vector v is equal to the number of processes the corresponding data element e_i is used in. This vector is called vector of absolute usage and serves as a quantitative measure of N_1. The second one is named vector of global simultaneous usage denoted by q whose elements are

$$q[i] = \sum_{j=1}^{n} \sum_{i=1}^{k} M[i, j]\, M[l, j]\, d_{lj}$$

where d_{lj} is 0 for $l = j$ and 1 otherwise. It is not difficult to realise that $q[i]$ is the number of other data elements used in conjunction with the data element e_i, so its value is related to the parameter N_2. The parameter N_3 is described by the use of third vector : vector of relative simultaneous usage t with k elements. In order to define this vector we must first establish auxiliary matrix R, $k \times k$ with elements

$$R[l,s] = \sum_{j=1}^{n} \frac{M[l,j]\,M[s,j]}{v[l]} \qquad [\%]$$

where $R[l,s] = 0$ for $l = s$ by definition. The element $R[l,s]$ is a measure of simultaneous occurrence of data elements e_l and e_s relative to the total number of occurrences of e_l. Based on R the elements of relative simultaneous usages are calculated as

$$t[i] = \frac{\sum_{s=1}^{n} a(i,s,z)}{q[i]}$$

The value $a(i,s,z)$ is equal to 1 if $R[i,s]$ is greater than parameter z and 0 otherwise. The parameter z is used as a border between large and small elements of R and it is usually taken to be around 70%.

Now, the axioms defining candidates for the three types of data elements turn into the following ones: a data element e_i is a candidate for:

- entity if $v[i]$ and $q[i]$ are small and $t[i]$ is small,
- strong entity attribute if $v[i]$ and $q[i]$ are small and $t[i]$ is large,
- weak entity attribute if $v[i]$, $q[i]$ are medium.

The final step in the (original) method is to decide upon the values of these three elements. This is done through the use of histograms (with adopted value of z), provided that the procedure is the same for all three of them. Let p be any of vectors v, q or t. Then the histogram is drawn with the x–axis representing the values of elements, and y–axis containing the number of elements that are equal to x, as seen on Figure 1. Such histogram tends to the normal distribution for vectors v and q, while for t it has two peaks. In any case, two points x_1 and x_2 on the x–axis are visually determined and values of p smaller than x_1 are taken to be small, the ones larger than x_2 to be large, and the rest is medium. The underlying argument is given in [1], so we shall not discuss it in detail.

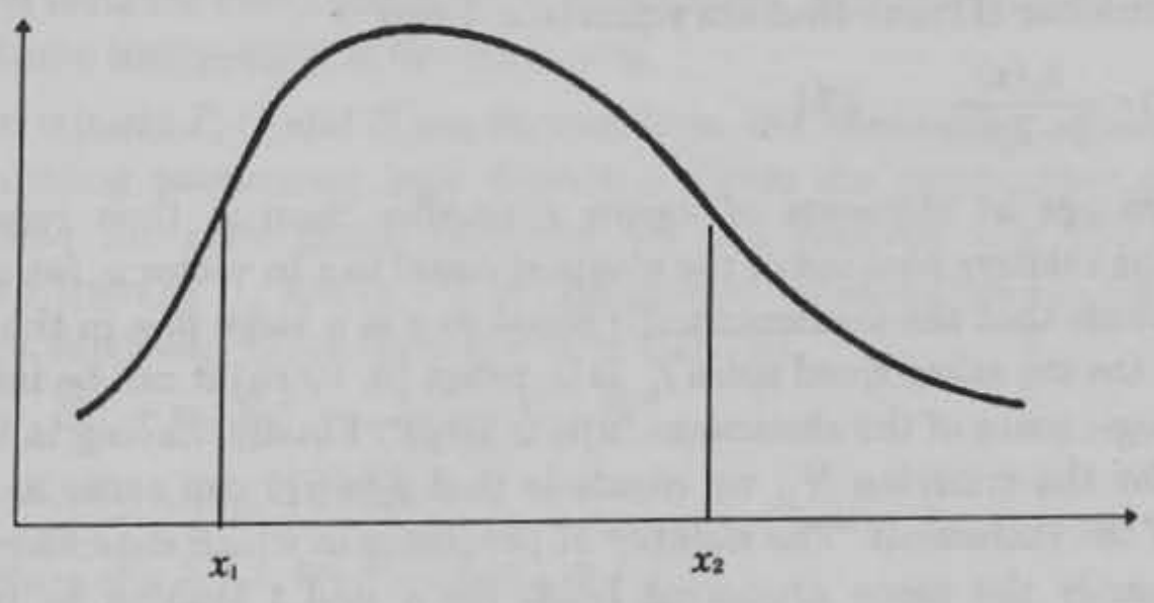

Figure 1.

The choice of x_1 and x_2 is made visually in such way that the majority of elements fall in the "medium" range.

The semi-subjective decision described has at least two considerable drawbacks. Firstly, the borders x_1 and x_2 are defined visually and even the sensitiveness of this method is impossible to predict, not to mention the potential errors. Second and, to our opinion worse, disadvantage is the lack of unique classification criterion because the decision on the type of data element is made on the basis of three criteria (v, q and t) which are not necessarily disjunctive. It would be obviously much better if the classification criterion was unique since it would eliminate or at least reduce the appearance of situations where the values of v, q and t are contradictory (such thing do happen).

2. QUANTIFICATION OF ELEMENTARY CRITERIA

To solve the two aforementioned problems, e.g. to minimize the level of subjectiveness present in the original method as well as to reduce the number of criteria, the method named *CLAS* (*Continuous Logic Attribute Synthesis*) is developed. It is generally based on continuous logic functions, and specifically on the method *LSP – Logic Scoring of Preference* introduced by J. Dujmović ([2], [6], described also in [3]), and which is intended for different purposes than data identification.

The main idea is to replace all Boolean logic statements in the model (such as "data element IS entity" or "Data element IS attribute" etc.) by continuous logic statements that are not only true or false, but can instead take all values between 100% meaning "true" and 0% meaning "false". Thus instead of stating for instance "e_i is an entity" we say "e_i is an entity with 80% confidence".

In order to introduce a background for continuous logic we must first transform the measures $v[i]$, $q[i]$ and $t[i]$, $i = 1, \ldots, k$, into elementary criteria ranging between 0% and 100%. Also the transformed values must in some way represent the position of particular vector element relative to the other elements. Let $b_v(x)$ be the number of v–values (elements of vector v) that are smaller than x where x is a free variable. Let $c_v(x)$ be the number of those that are equal to x. Then

$$f_v(x) = \frac{b_v(x)}{k - c_v(x)} \qquad [\%]$$

is the percentage of elements of vector v smaller than x, thus representing the measure of the relative position of the element equal to x in vector v. So, if $f_v(x)$ is large we may conclude that the v–element $v[i]$ equal to x is a large one in the environment defined by v. On the other hand since f_v is in range [0, 100%] it can be interpreted as a continuous logic score of the statement "$v[i]$ is large". Finally, having in mind that $v[i]$ is measure for the criterion N_1, we conclude that $f_v(v[i])$ can serve as a continuous logic score of the statement "The number of processes in which data element e_i is used is large". Exactly the same argument holds for q and t leading to two additional elementary criteria, f_q and f_t.

Using the same approach we define six additional elementary criteria: three of then g_v, g_q and g_t represent the logic score of statements "$v/q/t$ are small", and the rest h_v, h_q and h_t serve as a score of statements "$v/g/t$ are medium". It is easy to form their definition based on the relative position of the corresponding vector respectively as

$$g(x) = 100 - f(x) \quad [\%]$$

$$h(x) = 100 - 2\,\mathrm{abs}\,(f(x) - 50) \quad [\%]$$

where f is f_v for v–values, f_q for q–values and f_t for t–values.

Returning to the starting identification criteria we see that a data element e_i is (in the ideal case) an entity if

$$f_v(v[i]) = 100\% \qquad f_q(q[i]) = 100\% \qquad \text{and} \quad g_t(t[i]) = 100\%.$$

Consequently, it is an (ideal) strong entity attribute if

$$g_v(v[i]) = 100\% \qquad g_q(q[i]) = 100\% \qquad \text{and} \quad f_t(t[i]) = 100\%.$$

Finally, it is a weak entity attribute if

$$h_v(v[i]) = h_q(q[i]) = h_t(t[i]) = 100\%.$$

3. AGGREGATION OF CRITERIA AND TESTING

Looking at the results of previous paragraph we see that three global criteria may be constructed. The first is the one which would give a continuous logic value of statement "e_i is an entity" and it has to be constructed of the elementary criteria $f_v(v[i])$, $f_q(q[i])$ and $g_t(t[i])$. We denote it by $F(e_i)$. Next, the global continuous logic score of statement "e_i is a strong entity attribute" is evaluated through $g_v(v[i])$ and $g_q(q[i])$ and $f_t(t[i])$ and denoted by $G(e_i)$. The statement "e_i is a weak entity attribute" is evaluated through the global criterion $H(e_i)$ consisted of components $h_v(v[i])$, $h_q(q[i])$ and $h_t(t[i])$. To decide upon the type of data element e_i we compare the corresponding values of F, G and H and assign to it the type given by the statement with the largest score. It turns out, as seen later in this paragraph, that it is possible to establish logical tests for every global score individually, enabling thus to determine its level of correctness irrespective of the other two.

The global criteria F, G and H are formed from the elementary criteria by means of aggregation using continuous logic functions. Since the aggregation procedure is exactly the same for each global criterion we will present only the method for calculating $F(e_i)$ from f_v, f_q and g_t. The aggregation function used is one of general continuous logic functions which have the form ([2], [3])

$$E = (w_1 E_1^{\,r} + w_2 E_2^{\,r} + \ldots + w_m E_m^{\,r})^{1/r}$$

$$0\% \le E,\ E_1, \ldots, E_m \le 100\%$$

$$0 \le w_1, \ldots, w_m \le 1\,;\quad w_1 + \ldots + w_m = 1$$

r may be any real number,

where E_i are elementary criteria (in our case f_v , f_q and f_t), w_i are weight factors denoting the specific importance of every elementary criterion, and factor r determines the type of continuous logic function. Briefly, the continuous logic functions are divided into disjunctive type functions with $r > 1$, conjunctive type functions with $r < 1$ and the average ($r = 1$) which is of neither type. In order to make a distinction continuous logic functions have a prefix "quasi" (quasi–conjunction, quasi–disjunction). If any of components of quasi–disjunctive type is large the appropriate function tends also to be large (the larger r, the function is larger), while large values of quasi–conjunctive functions are obtained only if all components are simultaneously large. The most restrictive type of quasi–conjunction is the minimum function, whereas the least restrictive type of quasi–disjunction (the counterpart of former) is the maximum function.

When selecting the type of logic function to be used for aggregation we first notice that for the criterion F there is no reason to make any distinction between the relative importance of any component elementary criteria, so

$$w_1 = w_2 = w_3 = \frac{1}{3}$$

where w_1 is the weight factor of f_v , w_2 is the weight factor of f_q and w_3 is the same for g_t .

Deciding upon the type of the function is a slightly more complicated question. First of all, the function should be of the conjunctive type since it is necessary that all elementary criteria be satisfied simultaneously, the reason being that the data element type is determined by all three of them at the same time. This means that $r < 0$ meaning that the logic function type is conjunctive . For those functions there exists a wide spectrum of possibilities from the value near to arithmetic mean ($r = 1$) to the total conjunction with $r = -\infty$. Since there is no more information whatsoever which would be used to specify the exact value of r, we are forced to use some kind of average and this leads to the function called mean quasi–conjunction for which $r = -0.73$, [4], [5]. Thus, the aggregation of f_v , f_q and g_t is done by the use of criterion $F(e_i)$ of the form

$$F(e_i) = \left[\frac{f_v^{-0.73} + f_q^{-0.73} + f_t^{-0.73}}{3} \right]^{-1/0.73}$$

and $G(e_i)$ and $H(e_i)$ are calculated likewise.

Now $F(e_i)$, $G(e_i)$ and $H(e_i)$ are compared for every data element. It is an entity if the value of F is largest, a strong entity attribute if G is the largest one and weak entity attribute if the largest is H.

The continuous logic nature of F, G and H enables more powerful analysis then mere comparison. It is done through the use of logical formulae of the type (1) and an additional one which stands for continuous logic negation with an obvious definition

$$\neg \mathrm{E} = 100\% - \mathrm{E}$$

where $\neg$ E is a negation of E. This kind of negation is directly used for testing the opposite statements of the ones that define data element types.

Apart from that fairly simple case, every statement can be tested for logical correctness using the following reasoning : let e_i be the data element which is a candidate, let us say, for entity (the line of reasoning is identical for other two types). This means that $F(e_i) > \max(G(e_i), H(e_i))$. By using continuous logic it is possible to compute the logical level of statement "e_i is an entity and it is neither strong nor weak entity attribute". The corresponding expression is

$$C(F(e_i), \neg G(e_i), \neg H(e_i))$$

where C is one of conjunctive functions (the formula of type (1) with $r < 0$). Obviously, the same is possible for G and H.

4. CONCLUSIONS

The *CLAS* variant of *Attribute Synthesis method* gains its relative advantages by introducing continuous logic into the classification process. The main advantages if it are

- The level of subjectiveness is decreased, leaving only one parameter z (used in calculation of vector t) to be adopted provided that the range of its possible values is rather narrow. At the same time the visual determination of other variables based on histograms is completely eliminated.
- The number of independent criteria is reduced from three to one (maximum of F, G and H) eliminating thus the possibility of contradictory results.
- A powerful tool of continuous logic functions is available enabling various kinds of tests for statement correctness.

REFERENCES

[1] Teorey,T.J., and Fry,J.P., *Design of Database Structures*, Prentice-Hall, Englewood Cliffs, 1982.

[2] Dujmović,J.J., "Evaluation and Selection of Computer Systems", priv. comm.

[3] Malbaški,D. et al., "Comparative Study of Complex Systems by LSP Method", scientific project, Faculty of Technical Sciences, Novi Sad, 1980–1984.

[4] Malbaški,D., and Obradović,D., "A Contribution to the Data Elements Identification in Scientific Data Bases", in: *Proc. of 35 Yugoslav ETAN Conference*, Vol.9, Ohrid, (in Serbian), 1991.

[5] Malbaški,D., and Obradović,D., "Using CLAS method of Attribute Synthesis in the Environment of Structured Systems Analysis", in: *Proc. of 36 Yugoslav ETAN Conference*, Kopaonik, (in Serbian), 1992.

[6] Gilb,T., *Software Metrics*, Studentlitteratur, 1976.