Data Unification Algorithm for Representing Incomplete and Indefinite Information in the Medical Expert System

MANANA KHACHIDZE
Department of Computer Sciences
Tbilisi State University
University St. 2, Tbilisi 0143
GEORGIA
mkhachidze@tsu.ge

MAIA MIKELADZE
Department of Computing Techniques and Informatics
Tbilisi Institute of Economic Relations
Ketevan Tsamebuli str.30, Tbilisi, 0156
GEORGIA
mikeladzemaia@yahoo.com

Abstract: In the description of the patient, like any multi-system, we are confronting not only with abundance, but also with a large variety of factors (attributes, symptoms). In such a situation, the challenge becomes: how all this variety of features should be reduced to one type without information losing. Proposed algorithm helps to achieve this objective. The algorithm of data unification that allowing reducing all variety of exact and vague factors to one type without the loss of information is offered for that problem solution.

Key-Words: Data Presentation, Attribute System, Data Unification

1 Introduction
The diagnostics of the state of the complex multifactor system is one of the most topical problems in the human’s intellectual activity. That refers to every sphere of the human’s activity, such as medicine, economics, ecology, etc.

In the field of medical diagnostics the main problem is imprecision, incomplete and subjective data, which describe the condition of the patient. Added to this is non-rigidity medical knowledge. Despite this highly qualified doctor can to accurate diagnosis. In so doing, him helps not only the knowledge, but the intuition too.

The complexity of the problem diagnostics can cope only a scanty number of high qualified physicians-experts. as well as impossible to promptly transfer expert knowledge to other specialists, especially its unformalized part which being the result of the expert’s experience and intuition. in this case developing of models and methods for getting and presenting knowledge to aimed at the creation of expert systems is important.

2 Creation and Analysis Attribute System
In the design of expert systems, it is important to accurately find the methods of presenting data and knowledge. The application of the conceptual approach to the description of the states and the presentation of the knowledge as a procedure imitating the process of the notions forming by the natural intelligence is proved as a problem solution for showing up and formalization of the expert’s knowledge. Data used to medical data bases are distinguished by their diversity by presentation. When describing the state of the patient there arises a problem: how to reduce all the variety of features describing the state of the patient to one type not loosing the information, which at the same time would have allowed using a conceptual approach to the description of the state of the system.

3 Algorithm of Data Unification
3.1 Binary attribute.
Binary features are intended to describe the quality of symptoms that may have or not have occurred. Despite the variety of symptoms, to a question regarding those symptoms, "Do you grievance on ...?", patients given a standard response – either "yes" or "no". Therefore, we can all of these symptoms describe binary traits, the host 2 meanings: Consider what type of attribute needed for the description of any medical data and how they adapted to the conceptual approach. 1 - in the case of the presence of the symptom, 0 - in the absence of symptoms.

In applying the concept of a binary Ai comparison between the ground state vector \( \psi_i = (1) \) and \( \bar{\psi}_i = (0) \), and subsequent refinement of the concept is impossible.

3.2 Many Discrete Attributes
Many discrete attributes are intended to describe the symptoms that can take one of several values as numeric or character. The number of course and they
are mutually exclusive: for each patient takes only one sign of the importance and significance that one excludes the presence of other values.

The process of encoding such values and conducts them through a system of filters, to describe the conceptual approach [1]. We note that the values in the number of $2^m$ m levels allow a refinement of the concept.

If, for each patient attribute $A_i$ takes multiple values, and that one value does not exclude the existence of another value, then we are dealing either with any kind of data uncertainty (imprecise, fuzzy, inaccurate, incomplete), or poorly articulated sign.

Since conceptual method can be applied, unless the patient explicitly describes one of the state vector, $\psi_i$, or $\overline{\psi}_i$, it may be difficult.

In the particular case where the set of values taken by the $A_i$ attribute for a patient, is a subset of a vector, a clear description is possible.

Indeed, let $A_i = \{b_{i_1}, \ldots, b_{i_n}\}$ - The set of values, both a $A_i$ attribute for some patients, $\psi_i = (b_{i_1}, \ldots, b_{i_n})$ and $\overline{\psi}_i = (\overline{b}_{i_{n+1}}, \ldots, \overline{b}_{i_N})$, $N=2n=2^m$ - state vectors relevant of $A_i$.

Then if $A_i \subset \{b_{i_1}, \ldots, b_{i_n}\}$ the state is described by $\psi_i$ vector,

if $A_i \subset \{\overline{b}_{i_{n+1}}, \ldots, \overline{b}_{i_N}\}$ the state is described by $\overline{\psi}_i$ vector.

But in the general case, where $A_i'$ is a subset of any of the state vectors, one of the exits is a redefinition of the attribute. The original attribute is to be divided on such attributes, which would allow an unambiguous description of the vectors through the state. For example, each $b_{ij}$, $j=1,2,\ldots,N$ of $A_i$ attribute can be seen as a symptom that can have or not have a place, which compares the binary attribute

$B_{ij} = \begin{cases} 1 & \text{if presence } [a_{ij}, a^{i+1}_{ij}] \\ 0 & \text{if absence } [a_{ij}, a^{i+1}_{ij}] \end{cases}$ \quad j=1,2,\ldots,N$

with the state vectors $\psi_{ij}$ and $\overline{\psi}_{ij}$, $j=1,2,\ldots,N$.

Then, if the $A_i$ for a patient takes a lot of values $A_i = \{b_{ij}, \ldots, b_{ij}\}$, the state of the system of binary attribute features $B_{ij}$ ... $B_{iN}$ is described as $\psi_{ij} ... \overline{\psi}_{i_{k+1}} ... \overline{\psi}_{i_{k+1}}$.

### 3.3 Multivalued Continuous Attribute

Many attributes are continuing to describe the number of symptoms that can take values from some interval. In this case the state vector is not compared many elements, and numeric intervals. If the values of $A_i$ attribute is the interval $[a_i, b_i]$ a vector will be compared at level I:

$$\psi_i^{(1)} = \left[ a_i, \frac{a_i + b_i}{2} \right], \quad \overline{\psi}_i^{(1)} = \left[ \frac{a_i + b_i}{2}, b_i \right]$$

at level II:

$$\psi_i^{(2)} = \left[ a_i, \frac{3a_i + b_i}{4} \right], \quad \overline{\psi}_i^{(2)} = \left[ \frac{3a_i + b_i}{4}, b_i \right]$$

$$\psi_i^{(3)} = \left[ a_i + b_i, \frac{a_i + 3b_i}{4} \right], \quad \overline{\psi}_i^{(3)} = \left[ \frac{a_i + 3b_i}{4}, b_i \right], \text{etc.}$$

Intervals, compared vector at each level are halves of the interval - the vector of the state of the previous level, has been splitting. In contrast to the discrete multi attribute, a finite number of values which place a limit on the number of levels of refinement of the concept, in the case of multi lifelong attribute number of clarifying the concept, generally speaking, is not limited.

Otherwise, the area of $[a_i, b_i]$ multi continuous attribute should be split into "elementary" intervals $[a^{i^1}, a^{i^2}]$, $[a^{i^2}, a^{i^3}]$ ... $[a^{i^N}, b_i]$ In such a way that all the "interval" values taken on the set of patients, can be expressed as the union of a finite number of "elementary" intervals:

$$A_i = [a_i, a_i'] = [a_i^1, a_i^{i+1}] \cup [a_i^{i+1}, a_i^{i+2}] \cup \ldots [a_i^{i^N}, a_i']$$

Then each "basic" interval can be considered as a binary attribute

$$B_{ij} = \begin{cases} 1 & \text{if presence } [a_i^j, a_i^{j+1}] \\ 0 & \text{if absence } [a_i^j, a_i^{j+1}] \end{cases} \quad j=1,2,\ldots,N$$

each of which compared the state vectors $\psi_{ij}$ and $\overline{\psi}_{ij}$, $j=1,2,\ldots,N$.

If the $A_i$ attribute takes on the $A_i$ value, $A_i = [a_i^j, a_i'] = [a_i^1, a_i^{i+1}] \cup [a_i^{i+1}, a_i^{i+2}] \cup \ldots [a_i^{i^N}, a_i']$ a state of the patient in the system of binary $B_{ij}$ ... $B_{iN}$ be described as

$$\overline{\psi}_i^{(1)} \ldots \overline{\psi}_i^{(k+1)} \overline{\psi}_{ij} \ldots \overline{\psi}_{ij-1} \overline{\psi}_i^{(k+1)} \ldots \overline{\psi}_{i_{N}}$$

### 3.4 Linguistic Attributes

Linguistic attributes are intended to describe the fuzzy symptoms. Fuzziness of these symptoms are attributable either to the fact that their values are expressed as concepts, or inaccurate measurements,
incomplete and uncertain data. This may be a variable species \((X, T(X), U, G, M)\), where each term is clearly defined by a fuzzy set defined on the set of numbers \(U\). In the case where the value of linguistic variable can not be specified through the set of numbers, the variable is defined by listing its values - elements of term-sets: \((X, T(X))\).

Conceptual method can be applied to linguistic variables, as after coding their values we get the analogue of the discrete multi attribute, which is valid for all the comments in paragraph 3.2.

But in the case of the encoding of terms and in the case of partitioning the original feature in the binary traits of the information is lost, or rather, ignoring imprecision values. In the case of fuzzy a value of a single value only to some extent precludes the existence of other value may be as low as that in the presence of a discussion of approximate values, we would understand as a confirmation of the existence of another value.

In addition, once the value of the attribute expressed by linguistic terms, they may form a hierarchy: at the highest level of a more general notion, which is a generalization of its subsidiaries of vertices, and the existence of a notion does not exclude, but rather confirms the existence of its generalizations.

Consider this question in more detail.

### 3.6 Algorithm Description

As has been said, when describing the condition of the patient except for the abundance of attributes we are confronting with their great diversity. In such a situation, the challenge: how all this variety of attributes to reduce to a single mind, without loss of information that make it possible to build "concepts". And as has been shown, one of the exits in this situation is binarize attributes: each feature value \(b_j\), \(j = 1, 2, \ldots, N\) of \(A_i\) attribute is seen as a binary attribute.

\[
B_j = \begin{cases} 
1 \text{ if } b_j \text{ presence} \\ 
0 \text{ if } b_j \text{ absence} 
\end{cases} \quad j = 1, 2, \ldots, N 
\]

a patient condition in terms of \(A_i\) attribute in binary attributes would describe the "path"

\[
\psi_{i1} \ldots \psi_{iN}, \quad \sigma_j = \begin{cases} 
1 \text{ if } \psi_j \quad j = 1, \ldots, N \\
0 \text{ if } \bar{\psi}_j
\end{cases}
\]

where \(\psi_j\) and \(\bar{\psi}_j\) - state vector corresponding to the binary attribute \(B_j\), \(j = 1, \ldots, N\).

However, in the case of linguistic attributes, the values of which form a tree of concepts in which their children is to specify a more general notion of the parent peak, with a binarization loses of information \([3, 4]\).

Indeed, let \(A\) and \(C\) - the value of linguistic attribute \(L\), and \(C - A\) concept (Figure 1), while \(B_A\) and \(B_C\) - corresponding binary attributes. If the \(L\) attribute takes the value \(C\), then for binarization:

- \(B_C = 1\) (present value of \(C\))
- \(B_C = 0\) (absence of value \(A\)).

However, if there is a specification, it has a place its synthesis too,

- if \(L = C\), then \(B_C = 1\) and \(B_C = 1\).

Generally, the presence (absence) of concept and the presence (absence) of specification, the following dependencies:

1. If there is a specification, it has a place and its generalization.
2. If there is a general concept, it is nothing known about its specificity.
3. If there is no general concept, it is missing and its specification.
4. If there is no concrete definition, it is nothing known about its synthesis.

Based on the above relationships and notions of a tree can be drawn up to formal rules of true language attribute, allowing for the presence of any one tree tops to judge the value of (presence, absence or uncertainty) to the top of any concepts tree.

These are the formal rules \([2, 3, 4]\):

Let \(L\) - linguistic attribute, which values \(L_i\), \(i = 1, \ldots, N\) can be represented as a tree of concepts, with their children are mutually exclusive specification of a more general notion of the parent vertex (ie, each parent vertex is the OR - peak). Each value \(L_i\) of linguistic attribute put in the line triplet feature \(T_{L_i}\), which can take the following values:

\[
T_{L_i} = \begin{cases} 
0 \text{ (absence of values } L_i) \\
1 \text{ (presence of values } L_i) \\
2 \text{ (uncertainty) or the absence} \\
\text{presence or absence of values } L_i)
\end{cases}
\]

Let \(A, B, C, D, E\) - some of the top of the concepts tree (i.e. some values of the linguistic attribute \(L\)), and \(T_A, T_B, T_C, T_D, T_E\) - corresponding to these values triplet attribute, and know that \(L = C\), i.e. \(T_C = 1\) (Figure 1).

Then the values of triplet attribute that match the rest of the tops, will be determined as follows:

1. If the desired top - a level of \(C\) and with a common parent top, for example, the top \(A\) in Fig. 1, then if \(T_C = 1\), then \(T_D = 0\) (no value \(D\)).
2. If the desired top - the parent top of \(C\), for example, top \(A\) in Fig. 1, then if \(T_C = 1\), then \(T_d = 1\) (present value of \(A\)).
3. If the desired top - a subsidiary top of top C, for example, the top E in Fig. 1, then if TC = 1, TE = 2 (the uncertainty, we do not know value of E is present or absent).

4. In all other cases, such as top B in Fig. 1, we have if TC = 1, then TB = 0 (no value B).

Thus, in determining the values of triplet attributes of correlation between the presence (absence) and the concept of presence (absence) of the specification allow us to move the tree in a vertical direction, while the mutually exclusive tops provides the subsidiaries transitions in the horizontal direction.

4. In all other cases, such as top B in Fig. 1, we have if TC = 1, then TB = 2 (uncertainty).

Note that these rules are derived for a particular case - "clear" the tree of concepts. This tree compares linguistic attribute, expressed linguistic variable (L, T (L)), which is specified by listing the elements of thermo-set T(L). Each element of term-sets - it is an unclear concept, which is not described by fuzzy subsets of a numerical set. In this case, we can not calculate the degree to which a value of "rule out" other. The only thing we can do - to compare this variable, a tree of concepts in which the vertices are the a priori or absolutely mutually exclusive (OR-tree), or information about mutually exclusive is not defined (AND- tree).

In the general case, when the value of the linguistic attribute (L , T(L),U, G, M) defined as a fuzzy subset of a numeric set, then it is difficult to provide in the form of a concepts tree: due to ambiguities of L, i= 1, … N Each concept can be specification is not one, but several tops, and tops at the same level with the common parent top may or may not be "mutually exclusive" only to a certain extent. Here, we can not move in either a horizontal or vertical direction. In this case, the relationship between the presence (absence) of two concepts is determined by the degree of inclusion for fuzzy sets.

This is done as follows:

Let A and B - two meanings of linguistic attributes L, and TA and TB - corresponding to these values triplet signs.

\[ A = \{ \mu_A(u)/u \}, \quad u \in U \]
\[ B = \{ \mu_B(u)/u \}, \quad u \in U \]
- Fuzzy sets that describe the values A and B respectively, i.e. their presence.

\[ \overline{A} = \{ \mu_{\overline{A}}(u)/u \}, \quad \mu_{\overline{A}}(u) = 1 - \mu_A(u), \quad u \in U \]
\[ \overline{B} = \{ \mu_{\overline{B}}(u)/u \}, \quad \mu_{\overline{B}}(u) = 1 - \mu_B(u), \quad u \in U \]

Complement of fuzzy sets A and B, respectively, describing their absence.

Let L = A, i.e. TA = 1 and we want to determine the value of the attribute TB. Calculate

\[ \nu(A, B) \] – The degree of inclusion of fuzzy set A in fuzzy set of B

and

\[ \nu(A, \overline{B}) \] – The degree of inclusion of fuzzy set A in Fuzzy Sets \( \overline{B} \):

\[ \nu(A, B) = \mu_A(u) \rightarrow \mu_B(u) \],
\[ \mu_A(u) \rightarrow \mu_B(u) = (1 - \mu_A(u)) \vee \mu_B(u). \]

Then we set:

If \[ \nu(A, B) > \nu(A, \overline{B}) \], then to TB = 1 (present value

Fig. 1 concepts tree of Hierarchical attribute

Several different situations to when their children are not mutually exclusive and the presence of one of their children with a common parent vertex does not clarify the situation with their children (i.e., horizontal transitions are not defined, can move only in vertical direction). Accordingly, the rules triplication slightly changed [3, 4]:

Let know that L=C, i.e. TC = 1, then:

1. If the desired top - a level of C and with a common parent top, for example, the top D in Fig. 1, then if TC = 1, then TD = 2 (uncertainty);

2. If the desired top - the parent top of top C, for example, top A in Fig. 1, then if TC = 1, then TA = 1 (present value of A);

3. If the desired top - a subsidiary top of top C, for example, the tip E in Fig. 1, then if TC = 1, then TE = 2 (uncertainty);
of \( B \) 
If \( \nu(A, B) < \nu\left(A, \overline{B}\right) \), then \( T_B = 0 \) (absence value \( B \)) 
If \( \nu(A, B) < \nu\left(A, \overline{B}\right) \), then \( T_B = 2 \) (the uncertainty is unknown, has place \( B \) or \( \overline{B} \)). 

As a result, we obtain the pairwise correlation between the presence (absence) of any two values of the linguistic attribute \( L \), which will hold tripletization of this trait, without loss of information. Moreover, in case of insufficient information about the patient, we can in the presence of a symptom to judge the presence or absence of other symptoms, which would supplement the database and weaken the influence of incomplete data at the time of diagnosis.

4 Conclusion
In the description of the patient, like any multisytem, we are confronting not only with abundance, but also with a large variety of factors (attributes, symptoms). In such a situation, the challenge becomes: how all this variety of features should be reduced to one type without losing information, which the same would apply, a conceptual approach to the description of the system.

To achieve this, an algorithm is proposed to factors unification. This algorithm on the basis of correlation between the presence (absence) and the notion of generalization, as well as the degree of inclusion of one fuzzy set to another, allows to combine all the diversity of clear and unclear hierarchical features to a single mind, without losing information.

References: