Feature descriptions, linguistic variables,
comparison of descriptions of diverse types

Jakub ADAMCZAK[*], Piotr S. SZCZEPANIAK[**]

UNIFICATION OF FEATURE DESCRIPTION IN MEDICAL DATABASES

In this paper, the problem of unified analysis of data and descriptions of objects is discussed. Basic concept for measure of similarity between features described in diverse manner is presented as well as the method for unification of different types of description.

1. INTRODUCTION

Numerous databases and medical in particular, include data which is hard to standardise and process effectively. The information stored in linguistic form is completed by numerical data. There are also situations when numerical values and symbols are used to describe the same object or state. Consequently, the user must decide, whether he has to convert a part of data to the form acceptable by the computer system or he must limit the set of analysed data to the relevant amount and form. In this paper, an alternative solution is proposed.

To characterise patient condition, linguistic statements and numerical values are usually in use (e.g. blood pressure is stable; blood pressure is 84 [mmHg]). On the other hand, the state of health may be described by the comparison to its normal or expected state and represented by some value from the interval [0,1], e.g. almost normal – 0.8. In this sense, the degree to which some criterion is fulfilled can be interpreted in terms of the fuzzy sets theory [3, 5, 7]. Moreover, the fuzzy reasoning systems become applicable here.

Natural language is also frequently used for description of similarity of examined objects (or alternatively – of distance between them). For example, the feature called "general condition of patient" can be described by one of the linguistic terms from the following set

{„very bad", „bad", „satisfactory", „good", „very good"}. *(1)*

Note that each statement defining distance, like distance between "very bad" and "satisfactory" needs an expert explanation. It is obvious that experts may have different opinions; moreover, the object may be characterised by dozens (or hundreds) of features and in this case the uncertainty in precision of description grows.

[*] Dept. of Medical Informatics and Statistics, Medical University, pl. Hallera 1, 90-647 Łódź, Poland
[**] Institute of Computer Science, Technical University of Lodz, ul. Wólczańska 215, 93-005 Łódź, Poland

2. UNIFICATION OF FEATURE DESCRIPTIONS

Let us concentrate on the situation when the same feature of two or more objects is described by numeric and by linguistic values; in particular, in some records integer numbers are used while in others - linguistic descriptions. The problem lies in the comparison of these two objects, i.e. how to calculate the distance between linguistic and numeric descriptions. The natural way to deal with this problem is the fuzzy sets approach.

The approach is shown on an example when the "heart pulse" is considered. Medical definition of this feature is given in Table 1 to explain possible states of patient's heart condition. Let the three diverse descriptions are given, see descriptions (2)-(4),

$$\{deviated, normal\}, \tag{2}$$

$$\{180, 100, 90, 74\}, \tag{3}$$

$$\{40, 60, 73\}. \tag{4}$$

Table 1. Breakdown of heartbeats per minute with corresponding medical description [2]

value	**medical statement**
40 and lower	To low number of heartbeats per minute called *Bradykardia* – the patient's condition is critical
(40, 60]	Low number of heartbeats per minute Heart in bad condition
(60, 72] and (73, 90]	Heart in good condition
(90, 100)	High number of heartbeats per minute Heart in bad condition
[100, 180)	To high number of heartbeats per minute called *Tachykardia* – the patient's condition is critical
180 and higher	Extreme high number of heartbeats per minute the patient's condition is critical

Usually, "pulse" is represented by the number of heart beat per minute. Description sets (2)-(4) are constructed on the basis of values that are possible in reality (cf. Table 1). Here, two general forms of descriptions are considered: linguistic – (2), and numeric – (3), (4). Descriptions „deviated" and „normal", represent respectively bad and good activity of human heart in normal conditions. Moreover, the descriptions are arranged from "disfavourable" to "favourable" from medical point of view. Of course, the direction of the arrangement is free and it should not affect correctness of the unification proposed. There are some ways of the construction of the method for comparison of descriptions.

The normalisation of numeric values to the interval [0, 1] is straightforward. Let x_0 and x_k be the first and the last value of the considered records ($x_0, x_k \geq 0$; $x_0 \neq x_k$). Then for any x from the considered record of numeric description we have:

$$w(x) = \frac{x - x_0}{x_0 - x_k} \in [0,1] \tag{5}$$

In this way any numerical value describing the considered feature is represented by a number from range [0, 1]. For example, in the record (3) we have $x_0 = 180$, $x_k = 74$ (with $k = 4$), and the corresponding normalised representation is $w(x_0)=w(180)=0$, $w(x_1)=w(100)\approx 0.75$, $w(x_2)=w(90)\approx 0.85$, $w(x_4)=w(74)=1$.

On the other hand, linguistic descriptions can be represented as linguistic variables known from the theory of fuzzy sets. In description (3) one has two statements, and consequently two membership functions are sufficient for their representation; in Fig.1 triangular membership functions are used.

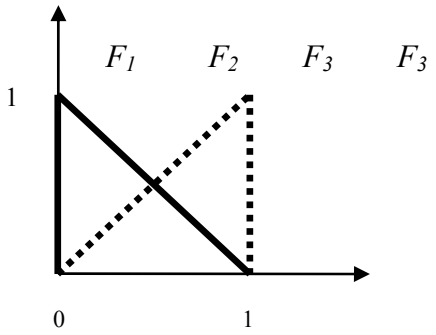

Fig. 1. Possible representation of (2)

Notice that the number of fuzzy sets related to linguistic variable *distance of i-th feature* depends on the used corresponding description. In the discussed example (2)-(4) there are three description sets with different number of values.

Define lz_j as the number of elements in the j-th description record, i.e. it is the value from the set

$$\{2,3,...,lz_{\max}\} \tag{6}$$

where $lz_{max} > 2$ and finite. For example, if (3) is considered then $j=2$ and $lz_2 = 4$.

In natural manner, the number of fuzzy sets is equal to the number of elements in the description record. In the simplest case, description consists of two linguistic or numeric values.

Define lb_i as a maximal value of all lz_j related to the i-th feature. In the example (2)-(4) of the "heart pulse", lb_i is 4. The number lb_i defines the final number of fuzzy sets used for comparison of linguistic and numeric values of all descriptions of i-th feature. General formula of the l-th membership function is

$$\mu_{F_l}(x;\sigma_i) = \max\left(\min\left(\frac{x-(l-2)\sigma_i}{\sigma_i}, \frac{l\sigma_i - x}{\sigma_i}\right), 0\right) \tag{7}$$

where

$$\sigma_i = \frac{1}{lb_i - 1} \tag{8}$$

Assuming that

$$a = (l-2)\sigma_i \quad b = (l-1)\sigma_i \quad c = l\sigma_i \tag{9}$$

one can define the triangular function as follows – cf. Fig.2

$$f(x;a,b,c) = \max\left(\min\left(\frac{x-a}{b-a},\frac{c-x}{c-b}\right),0\right) \tag{10}$$

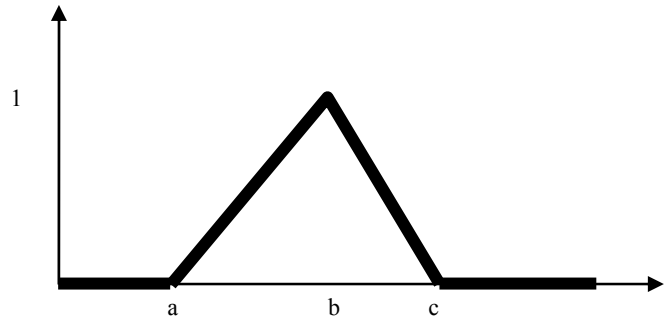

Fig. 2. Triangular function

Considering example of "heart pulse" one has $lb_i = 4$, and consequently four fuzzy sets for definition of the *distance of the i-th future* are used:

F_1 = identical,
F_2 = very similar,
F_3 = almost similar,
F_4 = different.

For comparison of the idea of the *distance of the i-th future* see [1]. Let us derive the appropriate membership functions for these fuzzy sets. The calculation of σ_i gives

$$\sigma_i = \frac{1}{4-1} \cong 0.333 \tag{11}$$

Finally, to each *distance of i-th future* corresponds one membership function defined as follows:

$$\mu_{F_1}(x) = \max\left(\min\left(\frac{x+0.333}{0.333},\frac{0.333-x}{0.333}\right),0\right) \tag{12}$$

$$\mu_{F_2}(x) = \max\left(\min\left(\frac{x}{0.333},\frac{0.666-x}{0.333}\right),0\right) \tag{13}$$

$$\mu_{F_3}(x) = \max\left(\min\left(\frac{x-0.333}{0.333},\frac{1-x}{0.333}\right),0\right) \tag{14}$$

$$\mu_{F_4}(x) = \max\left(\min\left(\frac{x-0.666}{0.333},\frac{1.333-x}{0.333}\right),0\right) \tag{15}$$

In Fig.3 this example is shown. To improve clarity of presentation the sets F_1 and F_3 are marked by the solid line while the others - by the dashed lines.

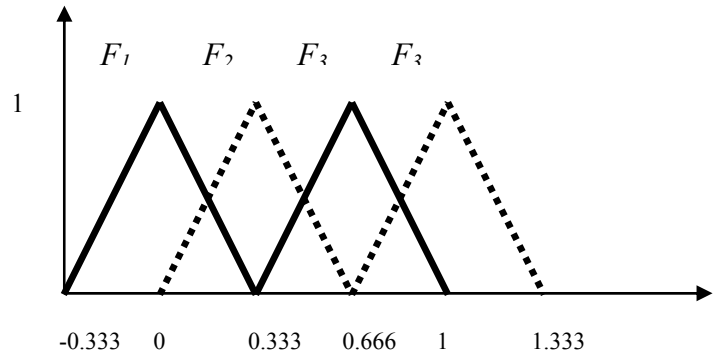

Fig. 3. Membership functions describing distance of the *i*-th future representing "heart pulse"

In this way, the comparison of descriptions given in diverse forms becomes easier and it can be seen as useful unification of feature descriptions.

3. SUMMARY

In this paper, the approach for automatic unification of descriptions existing in medical databases is proposed. Its application enables examination of the distance between linguistic and numeric values. Because of the variety of data and descriptions existing in medical databases any unification in development and interpretation of data is worth to consider. Having data in unified and comparable form, the system developer can concentrate on system's functionality and on the users' friendly interface. Similarly, the final user (medical staff) may apply any form of description which is convenient for him.

BIBLIOGRAPHY

[1] BUNKE H., FABREGAS X., KANDEL A. (2001): *Rule-Based Fuzzy Object Similarity*. Mathware & Soft Computing, 8, 113-128.
[2] HEROLD G. (2000): *Repetytorium dla studentów medycyny i lekarzy*. Wyd. Lekarskie PZWL, Kolonia, 2000.
[3] ŁĘSKI J., STRASZECKA E. (2003): Zbiory rozmyte i ich zastosowanie w diagnostyce medycznej. In the paper: Zajdel R., Kącki E., Szczepaniak P.S., Kurzyński M.: Kompendium informatyki medycznej. α–medica press, Bielsko-Biała.
[4] RUTKOWSKA D. (1997): *Inteligentne systemy obliczeniowe*. Akademicka Oficyna Wydawnicza PLJ, Warszawa.
[5] RUTKOWSKA D., PILIŃSKI M., RUTKOWSKI L. (1997): *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*. Wydawnictwo Naukowe PWN, Warszawa.
[6] WYDRZYŃSKI L.(1990): "INFRACTEST" – prognostyczny system ekspertowy; podręcznik użytkownika. Polgat, Katowice.
[7] SZCZEPANIAK P.S., LISBOA P.J., KACPRZYK J. (2000): *Fuzzy Systems in Medicine*. Physica Verlag, c/o Springer-Verlag, Heidelberg, New York.
[8] SZCZEPANIAK P.S. (2000): *Sieci neuronowe i logika rozmyta w medycynie – przegląd zastosowań*. In: *Biocybernetyka i inżynieria medyczna*. Ed. M.Nałęcz, tom 6 pt. *Sieci neuronowe*. Eds.:W.Duch, J.Korbicz, L.Rutkowski, R.Tadeusiewicz, 617-633. Akademicka Oficyna Wydawnicza EXIT, Warszawa.