A Fuzzy Set Theoretical Approach to the Automatic Generation of Absenteeism Analyses in Natural Language

N. du Bois
Mars 4
3402 JH IJsselstein
The Netherlands
n.dubois@planet.nl

M. De Cock, E. E. Kerre
Dept. of Applied Mathematics
and Computer Science
Ghent University
Krijgslaan 281 (S9)
B-9000 Gent, Belgium
Martine.DeCock@rug.ac.be

R. Babuška
Control Laboratory
Delft University of Technology
P.O. Box 5031
2600 GA Delft
The Netherlands
R.Babuska@et.tudelft.nl

Abstract

In this contribution a first important step is made towards the automatic generation of reports that describe (in natural language) absenteeism because of sickness in companies. Three parameters are examined: the sickness percentage, the absenteeism percentage (because of sickness), and the reintegration percentage. By means of fuzzy set techniques and an index of reintegration they are transformed into a linguistic description. The resulting technique has been successfully applied in the analysis of numerical absenteeism data of the year 2000 in about 20 divisions of a big company in the Netherlands.

Keywords: sickness absenteeism, reintegration, linguistic term, linguistic modifier, linguistic interpretation, fuzzy set, fuzzy clustering, health care, company strategy

1 Introduction

In the Netherlands an employer pays the salary of an ill employee during the first year of absence. Both employer and employee are responsible for a quick recovery and resumption of work (reintegration). All employers are obliged to join in with a company health service, who provides them with an absenteeism report about the passed year every January. Such kind of reports contain a lot of numerical data: sickness percentage, absenteeism percentage (because of sickness), reintegration percentage, the frequency of reporting sick, etc. If the employer wants an analysis of those figures, e.g. by contract with the health service, this analysis is done by occupational physicians, by hand. For them this is an extra workload in one of the most busy months of the year with respect to the high absenteeism because of sickness — January is part of the winter season in the Netherlands — and the resulting crowded office hours in that month. Still the analyses have to be done quickly because the company strategies for the coming year depend on them.

It is clear that an automatic generation of absenteeism analyses would save a lot of time and effort. In this contribution we present an approach towards this. An absenteeism analysis consists mainly of natural language expressions. To model these we use fuzzy sets. More specifically, we will show how membership functions for linguistic terms suitable for the description of the percentages above can be constructed, and how they can be used for the interpretation of the numerical data into natural language sentences. In the case of sickness percentage (SP) and of absenteeism percentage (AP) this is quite straightforward. For the reintegration however we have to rely on another measure — called the reintegration index (RI) — which is based on the two former percentages, and a fortiori also on the reintegration percentage (RP).

2 The Need for a Reintegration Index

The percentage of employees of a company that are sick is referred to as the sickness percentage SP. Not all of those employees will be absent from work. Usually a percentage of the employees is sick and absent (the absenteeism percentage AP), while another percentage is sick but still goes to work (the reintegration percentage RP). I.e.

$$SP = AP + RP$$

It is the task of the manager to keep the AP as low as possible. This can not only be done by keeping the SP low, but also by controlling the RP. E.g. a postman who broke a leg will not be able to deliver letters during several weeks but he can be reintegrated in a job at the post office. A general rule for the manager is thus to keep the RP "high".

However the absolute value of the reintegration percentage alone does not give enough information to be able to interprete the degree of reintegration correctly. When the sickness percentage is 0.001, then the reintegration percentage of course is smaller than or equal to 0.001. But despite of the fact that the RP is very close to its smallest possible value (namely 0) this is obviously not a problematic situation with regard to reintegration. When the sickness percentage is 18, then a reintegration percentage of 1 can be interpreted as "low". When on the other hand the sickness percentage is 1.5, then a reintegration percentage of 1 is"high". Clearly the interpretation of degree of reintegration depends not only on the RP. Therefore another parameter for the degree of reintegration has to be introduced. A first step to achieve a better interpretation is to take the sickness percentage as a comparison measure. This can be done by using the reintegration ratio RR defined by

$$RR = RP/SP$$

For a more refined measure the AP can be involved as well, resulting in the reintegration index RI [8]:

$$RI = RP/(SP * AP)$$

Discussion about this issue is not ended yet. However in the research reported in this paper the reintegration index was used to give an interpretation of the degree of integration.

3 Construction of the Membership Functions

In fuzzy-set-theoretical systems linguistic terms are modelled by means of fuzzy sets. A fuzzy set on a universe X is characterized by its membership function, i.e. a mapping from X to $[0, 1]$. The class of all fuzzy sets on X is often denoted $\mathcal{F}(X)$.

Typically the construction of suitable membership functions is one of the most difficult tasks when building an application. The use of linguistic modifiers — such as more or less and very — facilitates this job since it allows for the automatic deduction of new membership functions from the existing ones using fuzzy modifiers (as we will discuss further on). Nevertheless first we have to come up with membership functions for at least some basic terms. In our case the linguistic trichotomy [14] will be formed by low, average and high.

Basic terms. At our disposal is an amount of historical data regarding SP, AP, and the derived RP and RI, as well as the opinion of an expert. This calls for the combination of a data-driven and a knowledge-based approach. Concerning the knowledge-driven part, as only one expert is involved, we can not consider techniques accumulating the opinion of a group of persons as was done e.g. in [15]. Therefore the role of the expert will be the evaluation and possibly the correction of the membership functions generated in a data-driven manner. For this purpose we have chosen to use a fuzzy clustering algorithm (see [1]).

To find suitable membership functions for the linguistic terms low, average and high for the variable SP, all SP values of divisions of the Dutch postal service of 1994 until 1996 were

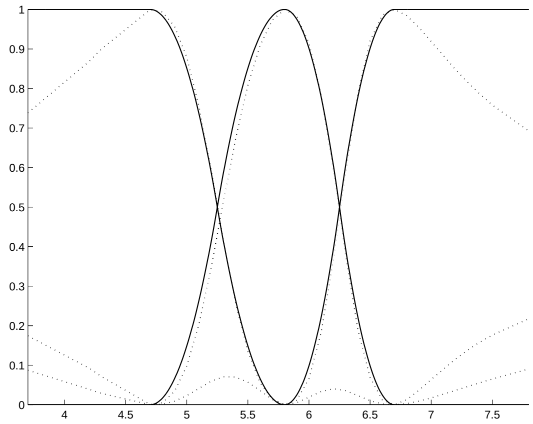

Figure 1: SP values of 1994 until 1996 divided into 3 fuzzy clusters (dashed lines) and their approximations (solid lines) .

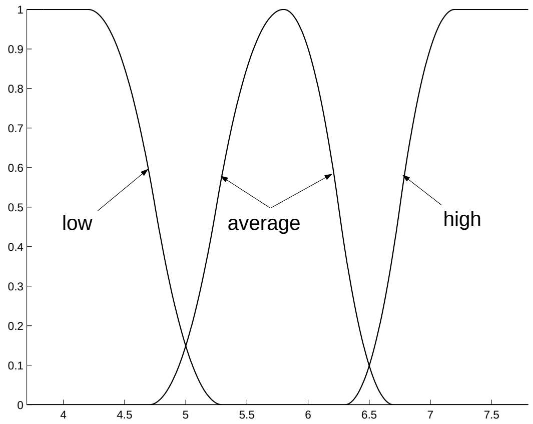

Figure 2: Membership functions for the basic terms describing the SP .

divided into three fuzzy clusters which were then approximated by S- and Π-membership functions (see Figure 1 for the resulting membership functions). We recall that the S- and Π-membership functions are characterized by respectively three and four real parameters and that they are defined by, for all x in $\mathbb{R}$:

$$S(\alpha,\beta,\gamma,x)=\begin{cases}0, & x\leq\alpha\\ \frac{2(x-\alpha)^2}{(\gamma-\alpha)^2}, & \alpha\leq x\leq\beta\\ 1-\frac{2(x-\gamma)^2}{(\gamma-\alpha)^2}, & \beta\leq x\leq\gamma\\ 1, & \gamma\leq x\end{cases}$$

and $\Pi(\alpha,\beta,\gamma,\delta,x)$

$$=\begin{cases}S(\alpha,(\alpha+\beta)/2,\beta,x), & x\leq\beta\\ 1, & \beta\leq x\leq\gamma\\ 1-S(\gamma,(\gamma+\delta)/2,\delta,x), & \gamma\leq x\end{cases}$$

It is assumed that $\alpha\leq\beta\leq\gamma\leq\delta$. S-membership functions are commonly used to represent increasing notions such as high, while the complement of an S-function is useful to model a decreasing notion such as low.

The technique described above can also be applied to obtain membership functions for the same basic terms for the variables AP, RP and RI for the overall figures, as well as for AP for the level of the subsets of the overall diagnoses: diseases of the locomotor system, psychological diseases, diseases of the heart and bloodvessels, diseases of the airways, accidents and a remainder group. Furthermore the frequency of reporting sick FS can be analyzed in the same way.

Modified terms. Typically in an application 7 ($\pm$ 2) linguistic terms are used, because this magic number corresponds to the number of distinctions a human being can perceive (see [13]). Therefore we can still add between 2 and 6 linguistic terms. To this end we will choose terms that are generated by applying linguistic modifiers to the basic terms we are already considering.

Adding new terms allows for more expressivity and diversity but may for this reason also cause a slight change in the meaning of the basic terms. Indeed if we also consider terms such as very high and more or less high, the size of the "area" of percentages originally "covered" by high will decrease. To deal with this phenomenon, in dialogue with the expert, we have decided to shift the left and the right membership function of Figure 1 a bit away from the center to allow for a more detailed range in between. This resulted in the membership functions depicted in Figure 2.

Fuzzy modifiers (also called *fuzzy hedges*) are mappings that transform a membership function into another. They are best-known as tools for the representation of linguistic modifiers such as very, more or less, at least, roughly, etc. During the last 3 decades to this purpose many fuzzy modifiers were suggested in the literature [12]. Although already useful, they have important shortcomings, which are due to the fact that they are designed simply to perform a technical transformation. Recently two new approaches were developed in which the representation of linguistic modifiers is endowed with an inherent semantics: the horizon approach [14] and the framework of fuzzy modifiers based on fuzzy relations [2, 5].

The strength of the latter approach is that in determining the degree to which y is more or less t and the degree to which y is very t, the context is taken into account, namely the fuzzy set of all objects resembling y. Resemblance is modelled by means of a fuzzy relation R on X, which is a fuzzy set on $X \times X$. For y in X the R-foreset of y is the fuzzy set Ry on X defined by $(Ry)(x) = R(x, y)$, for all x in X. If R is a resemblance relation on X, i.e. R is a fuzzy relation on X such that for all x and y in X, $R(x, y)$ is the degree to which x and y resemble each other, then Ry is the fuzzy set of objects resembling y.

The general idea is that an object y can be called more or less t if it resembles an object that can be called t. Likewise an object y can be called very t if every object it resembles can be called t. Assuming that the linguistic term t is modelled by means of the fuzzy set A, in the first case we need to represent the intersection of Ry and A for which we will use a triangular norm (a generalization of the boolean "and" operator to $[0, 1]$). In the second case we have to study the inclusion of Ry in A; to this end we will need another tool from fuzzy logic, namely an implicator.

A *triangular norm* (shortly t-norm) $\mathcal{T}$ is an increasing, associative and commutative $[0, 1]^2 \to [0, 1]$-mapping satisfying the boundary condition $\mathcal{T}(1, x) = x$, for all x in $[0, 1]$. An *implicator* $\mathcal{I}$ is a $[0, 1]^2 \to [0, 1]$-mapping with decreasing first partial mappings $\mathcal{I}(., x)$ and increasing second partial mappings $\mathcal{I}(x, .)$ that satisfies $\mathcal{I}(1, x) = x$, for all x in $[0, 1]$.

For A and B two fuzzy sets on X the **degree of inclusion** and **the degree of overlap** are defined by:

$$\mathsf{INCL}(A, B) = \inf_{x \in X} \mathcal{I}(A(x), B(x))$$

$$\mathsf{OVERL}(A, B) = \sup_{x \in X} \mathcal{T}(A(x), B(x))$$

Using these notions the following representation can be constructed:

$$(\text{more or less } A)(y) = \mathsf{OVERL}(Ry, A))$$

$$(\text{very } A)(y) = \mathsf{INCL}(Ry, A)$$

These representations correspond to the direct and the superdirect image [11] of A under R, i.e.

$$\text{more or less } A = R(A) \text{ and very } A = R^{\triangleright}(A)$$

These images respect all kinds of mathematical properties [4] which can be interpreted for linguistic terms [3]. Figure 3 depicts fuzzy sets for the linguistic terms for the SP. The membership functions for the modified terms were derived from those of Figure 1 using fuzzy modifiers based on the fuzzy relation

$$R(x, y) = \Pi(x - 1, x - 0.2, x + 0.2, x + 1, y)$$

Note that R is not a fuzzy equality nor a fuzzy equivalence relation although it clearly models approximate equality from an intuitive point of view (see also [6], [7]). If the difference between two percentages x and y is smaller than or equal to 0.2, then x and y are considered to be approximately equal to degree 1. If the difference is greater than or equal to 1, they are not considered to be approximately equal (i.e. they are considered to be approximately equal to degree 0). In between there is a gray area in which they are considered to be approximately equal to some degree between 0 and 1.

Furthermore we used the Łukasiewicz t-norm and implicator which are defined by

$$\mathcal{T}_W(x, y) = \max(x + y - 1, 0)$$

$$\mathcal{I}_W(x, y) = \min(1 - x + y, 1)$$

for all x and y in $[0, 1]$.

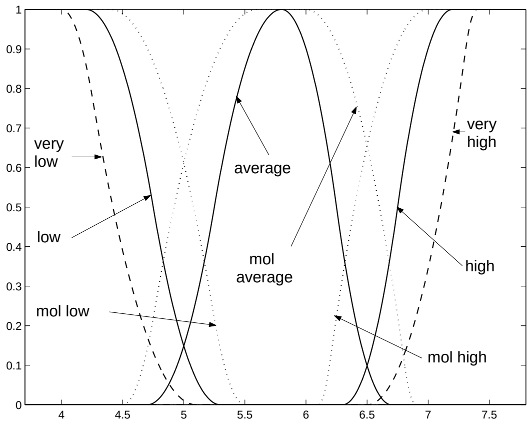

Figure 3: Inclusive interpretation: membership functions for (from the left to the right) very low, low, more or less low, more or less average, average, more or less high, high, very high.

In Figure 3 it is clear that the application of fuzzy modifiers based on R can result in a change of the kernel and the support, which makes them intuitively more plausible than the powering modifiers [16]. In the same figure it is shown that these fuzzy modifiers based on R can be applied in a uniform way to an increasing, a decreasing and a partially increasing and decreasing membership function, which is not possible for shifting modifiers [10, 13].

Combining fuzzy clustering, expert knowledge and fuzzy modifiers, we have established the membership functions of 8 linguistic terms for each variable under consideration. Although, using the superdirect image, it was possible to generate a fuzzy set for very average as well, we explicitly chose not to do this since in the literature not everybody agrees that very should be applied to a medium term such as average (see [14]). To some people very average might even have a negative connotation. Taken into account our goal — namely building a system that generates natural language absenteeism reports for occupational physicians and managers — we avoid terms that might be ambiguous or that seem not natural to some people.

4 Chosing the best term

To perform the interpretation of numerical data into natural language expressions, we have to chose the most suitable term for every crisp numerical input, i.e. the most suitable fuzzy set. To achieve this we have explored two techniques. Let us start by noting that the fuzzy sets under consideration can be divided in three groups, namely $\mathcal{A} = \{A_1, A_2, A_3\}$, $\mathcal{B} = \{B_1, B_2\}$ and $\mathcal{C} = \{C_1, C_2, C_3\}$ with A_1, A_2, A_3, B_1, B_2 and C_1, C_2, C_3 respectively corresponding to very low, low, more or less low, more or less average, average, more or less high, high, and very high. Within every group the ordering $\subseteq$ defined by, for all A and B in $\mathcal{F}(X)$,

$$A \subseteq B \text{ iff } (\forall x \in X)(A(x) \leq B(x))$$

is total, because the fuzzy relation R is reflexive [4].

In [9] it is suggested to chose a threshold c in $[0, 1]$, meaning that all membership degrees greater than c are considered as “maximal”. Within a group the most suitable fuzzy set for a crisp input x is chosen as the smallest fuzzy set A (w.r.t. $\subseteq$) such that $A(x) \geq c$. If a most suitable term arises from more than 1 group at the same time, then either the threshold should be increased, or a preference relation among the groups should be defined. If no suitable term arises, then one may consider decreasing the threshold.

Another option is to switch from the inclusive interpretation to the non-inclusive interpretation [12]. Until now we have assumed that the ordering $\subseteq$ is total in every group of terms (in fact we have modelled the fuzzy sets in this way). We have for instance made sure that the membership function for very low is *included* in the membership function for low. The underlying semantics is that every percentage that is very low is also low. In this interpretation the membership degree of x in a fuzzy set A clearly corresponds to the degree to which x *satisfies* the term modelled by A: indeed the degree to which a percentage is low is always greater than or equal to the degree to which it is very low. To distinguish this interpretation more clearly, the membership functions could be labelled with a preceding at least.

In the non-inclusive interpretation however the fuzzy sets within a group do not necessarily denote subsets nor supersets of each other, but different, possibly overlapping categories. In this interpretation the membership degree of x in A corresponds to the degree to which x is *representative* for the term modelled by A. Or vice versa: if $A(x)$ is the highest membership degree of x in all the fuzzy sets, then among these, A models the most representative term for x.

In [15] it is briefly explained how membership functions for the non-inclusive interpretation can be derived from those of the inclusive interpretation. The method is described for membership functions that are increasing in the inclusive interpretation (namely the modified linguistic terms of the increasing notion annoyed), but we can extend it to monotonic membership functions (increasing *and* decreasing membership functions), i.e. those of group $\mathcal{A}$ and group $\mathcal{C}$. The non-inclusive counterpart A' of a fuzzy set A belonging to either one of these groups can be established using the greatest fuzzy set B of the same group as A that is still included in A in the following way:

$$A' = A \cap co(B)$$

in which $\cap$ and co are the usual fuzzy intersection and fuzzy complement respectively defined by means of the minimum t-norm and the standard negator $\mathcal{N}(x) = 1 - x$ (needless to say generalization of these fuzzy logical connectives is possible). The non-inclusive representation of low for example corresponds to the inclusive representation of low but not very low. If for some fuzzy set A the fuzzy set B does not exist, then $A' = A$.

Regarding group $\mathcal{B}$ we chose to model average in the non-inclusive interpretation as average but not more or less low and not more or less high i.e.

$$B_2' = B_2 \cap co(A_3) \cap co(C_1)$$

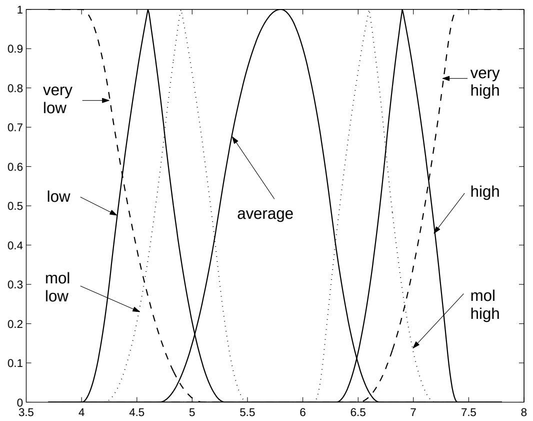

Figure 4: Non-inclusive interpretation: membership functions for (from the left to the right) very low, low, more or less low, average, more or less high, high, very high.

We omitted more or less average from the list, avoiding the open question whether in the non-inclusive interpretation more or less average should still be represented by a superset of average or rather by 2 “bumps”, one to the left and one to the right of the membership function for average. The resulting normalised membership functions are depicted in Figure 4.

In this way about 20 analyses of different companies were made in much less time than usually required. In these analyses an interpretation was made about SP, AP, RP for the overall figures and AP for the level of the subsets of the overall diagnoses. The results were very promising and were seemingly very much the same as those made by experts. They were even likely to be more consistent.

5 Conclusion and future work

The use of fuzzy set theory offers a way to automate analyses of absenteeism. Suitable membership functions for linguistic terms describing the variables can be derived automatically from historical data, relying on fuzzy clustering (for the basic terms) and fuzzy modifiers (for the modified terms). This makes the system very flexible and easy to handle. Using the generated fuzzy sets, new numerical data were transformed into natural language expressions. The resulting absenteeism analyses were already very satisfactory and could be used in every day practice.

Further work may involve the implementation of an approximate reasoning scheme that allows for the interpretation of reintegration solely from the SP and the AP without the use of the RI. It would be interesting to compare the results arising from such a system against the results obtained with the method presented in this paper. The ultimate goal is the developement of a continuous monitoring system on the computer that does not only interprete numerical data regarding SP, AP, FS, ... but also guides (or even replaces) the manager in making appropriate decisions such as *"Continue present approach." "Take on special program to prevent back, neck and shoulder complaints." "Focus highly on reintegration during the first year of illness."*

Acknowledgements

M. De Cock is supported by the Fund for Scientific Research - Flanders (FWO). Part of the research reported on in this paper was conducted during a stay of N. du Bois and M. De Cock at Delft University of Technology; both co-authors would like to thank the people of the Control Laboratory for their kind hospitality.

References

[1] R. Babuška (1998). Fuzzy Modeling for Control. Kluwer Academic Publishers, Boston. A fuzzy identification toolbox for MATLAB which implements some of the methods described in the book is available.

[2] M. De Cock, E. E. Kerre (2000). A New Class of Fuzzy Modifiers. Proceedings of ISMVL2000, IEEE Computer Society, pp. 121–126.

[3] M. De Cock, A. Radzikowska, E. E. Kerre (2000). Modelling Linguistic Modifiers Using Fuzzy-Rough Structures. Proceedings of IPMU2000, Volume III, pp. 1735–1742.

[4] M. De Cock, M. Nachtegael, E. E. Kerre (2000). Images under Fuzzy relations: A Master-Key to Fuzzy Applications. Intelligent Techniques and Soft Computing in Nuclear Science and Engineering, Proceedings of FLINS 2000. Edited by D. Ruan, H. A. Abderrahim, P. D'hondt, E. E. Kerre. World Scientific, pp. 47–54.

[5] M. De Cock, E. E. Kerre (2001). Fuzzy Modifiers Based on Fuzzy Relations. Accepted for Information Sciences.

[6] M. De Cock, E. E. Kerre (2001). Approximate Equality is no Fuzzy Equality. In: Proceedings of EUSFLAT2001, p. 369-371.

[7] M. De Cock, E. E. Kerre (2001). On (un)suitable Fuzzy Relations to Model Approximate Equality. Accepted for: Fuzzy Sets and Systems.

[8] N. du Bois (1997). About the Reintegration Efforts of PTT Post and the Introduction of Fuzzy Logic in its Analysis, (in Dutch) research paper for Social Medical Professional Education, Catholic University of Nijmegen.

[9] A. Dvořák, V. Novák (2001). Fuzzy Logic Deduction with Crisp Observations. Submitted to Soft Computing (private communication).

[10] H. Hellendoorn (1990). Reasoning with Fuzzy Logic. Ph. D. Thesis, T.U. Delft.

[11] E. E. Kerre (1993). Introduction to the Basic Principles of Fuzzy Set Theory and Some of its Applications. Communication and Cognition, Gent.

[12] E. E. Kerre, M. De Cock (1999). Linguistic Modifiers: an overview. In: Fuzzy Logic and Soft Computing (G. Chen, M. Ying, K.-Y. Cai, eds.), Kluwer Academic Publishers, pp. 69–85.

[13] G. Lakoff (1973). Hedges: a Study in Meaning Criteria and the Logic of Fuzzy Concepts. Journal of Philosophical Logic, 2, pp. 458–508.

[14] V. Novák, I. Perfilieva, J. Močkoř (1999). Mathematical Principles of Fuzzy Logic. Kluwer Academic Publishers, Boston/Dordrecht/London.

[15] A. Verkeyn, M. De Cock, D. Botteldooren, E. E. Kerre (2001). Generating Membership Functions for a Noise Annoyance Model from Experimental Data. To appear in: Soft Computing in Measurement and Information Acquisition (L. Reznik, V. Kreinovich, eds.), Studies in Fuzziness and Soft Computing, Springer-Verlag.

[16] L. A. Zadeh (1972). A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges. Journal of Cybernetics, 2, 3, pp. 4–34.