# LINGUISTIC VARIABLES DETERMINATION USING FUZZY CLUSTERING

**Antonio Flores-Sintas**[1]
Dpto. de Nuevas Tecnologías
Caja Ahorros de Murcia
Gran Vía,23
30005-Murcia
e-mail: aflores@cajamurcia.es

**José M. Cadenas**
Dpto. Informática, Inteligencia Artificial y Electrónica
Facultad de Informática
Universidad de Murcia
30071-Espinardo.Murcia
email: jcadenas@dif.um.es

**Fernando Martín**
email: fmartin@dif.um.es

## ABSTRACT

The fuzzy sets defining the linguistic variable values can be seen as a fuzzy partition of the linguistic variable. The membership functions obtained using fuzzy clustering algorithms are defined with respect to the group prototypes, and they cannot be used to define the linguistic variable values. We introduce several criteria to pass from the clustering membership functions to the linguistic variable value membership functions. The data have been taken from the financial sector.

**Keywords:** membership functions, cluster analysis, fuzzy c-means.

## 1. INTRODUCTION.

"The clients must guide the actions of the financial companies and these must adjust the offer of products and services to satisfy the needs of the clients. The marketing policy must be founded on strategic lines of action just as to classify the clients, to make blocks of products, to personalise the financial products and to offer products and services of quality. However, it is not easy to classify the clients, and it is more difficult to do it if we want classify clients belonging to the home economies sector because this is a wide and heterogeneous set. The classification must be pointed toward the market. For this reason, among the possible criteria to be used, it is necessary give a greater relevance to the advantages searched by the clients when they are related with the financial companies, and to the products and services used by the clients".(translated from [2])

The financial sector is highly computerised and it is a direct sales sector. As a consequence, the information about the consumption of the clients is very high. However, the number of products and services offered is, as any direct sales sector, much larger than the average number of products and services consumed by client. This fact makes difficult the application of conventional techniques of classification in the features global space (products and services) [1]. On the other hand, the bank clerks daily realise a classification of their clients using linguistic variables of the type: he has a medium debit balance, a low loan, a high investment fund, etc. It seems reasonable to use the same criterion and to realise one-dimensional classifications to establish the profile of the clients. In [6] we have realised an approximation to this problem, and this contribution is a sequel of [6].

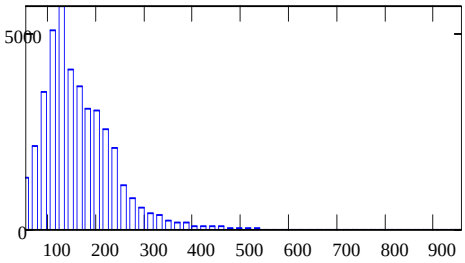


Figure 1. Number of clients/feature value

To determine the profile of a client we can use any countable feature providing information about the client. In this sense as example, the interest payment to a client can be used as a feature although it is not a product or a service offered but a financial cost. We have detected two features types distributed as figures 1 and 2. The salaries payment or the charges of the home insurances are as figure 1. The interest payment or the charges of the life insurances are as figure 2. We can see, for both cases, that there are not natural groups and that the distributions are very not homogeneous (the data are taken from Caja de Ahorros de Murcia). Referring to the first problem, we are not interested in detecting natural groups but in dividing a one-dimensional sample in a number of groups adjusting them as to the values of a linguistic variable (very-low, low, medium, high, very-high). In the

[1] Corresponding author

contribution we will use five groups corresponding to the five previous linguistic values. Referring to the second problem we cannot use the Euclidean distance because the distributions are very not homogeneous ([3],[4],[5]). We will use the Mahalanobis distance and the Fuzzy C-Means algorithm (FCM), modified in [4] which is based on [3], to find a good partition. However, the grades of membership found by the FCM algorithm are referred to the detected prototypes and they cannot be used as grades of membership to the values of the linguistic variable. In this paper, we introduce several criteria to pass from the membership functions found by the FCM algorithm to the membership functions defining the values of the linguistic variable.

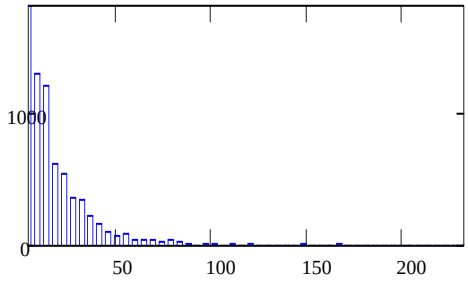


Figure 2. Number of clients/feature value

The FCM algorithm obtains the prototypes by Picard iterations from any hard partition, minimising a within-group sum of squared errors objective function. The FCM algorithm modified in [4] (it can also be in [6]) differences between the grades of membership and the probabilities of membership. In this paper we only are interested in the grades of membership, which are given by the following expression:

$$\mu_{xk} = \sqrt{\left|C_k^{-1}\right|} \Big/ \left(1 + d_{xk}^2\right)^{3/2}$$

$\mu_{xk}$ being the grade of membership of the $x$-pattern to the $k$-group; $C_k$ being the $k$-group fuzzy covariance matrix, and $d_{xk}^2 = \left(x - v_k\right)^T C_k^{-1} \left(x - v_k\right)$ being the Mahalanobis distance between the $x$-pattern and the $v_k$-prototype. Notice that the range of the membership function is different from the unit interval. This is not contradictory with the fuzzy set theory [7].

Figures 3 and 4 represent the membership functions obtained by using the FCM algorithm corresponding to the samples whose distributions are shown in figures 1 and 2.

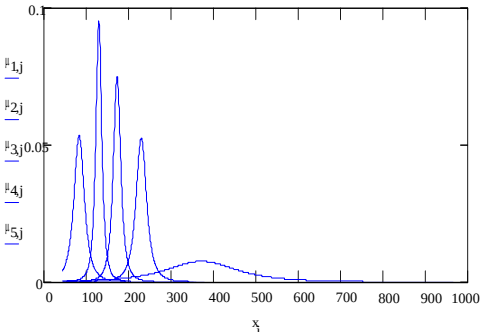


Figure 3. Membership functions (FCM) corresponding to figure 1.

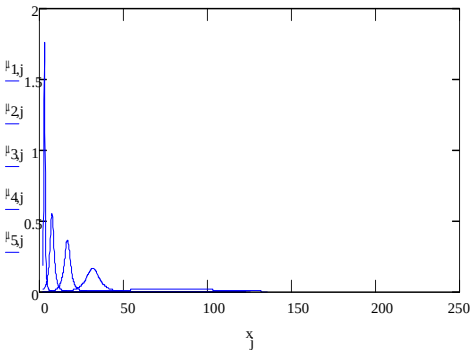


Figure 4. Membership functions (FCM) corresponding to figure 2.

## 2. MEMBERSHIP FUNCTIONS TO THE LINGUISTIC VARIABLE VALUES.

To distinguish between the membership functions calculated by the FCM algorithm and the membership functions of the linguistic variable values, we are going to represent by the symbol $\eta$ the second membership functions, maintaining the symbol $\mu$ for the first ones. $\eta_{xk}$ will represent the grade of membership of the $x$-pattern of the linguistic variable value of the $k$-group, being the 1-group=very-low, 2-group=low, etc., or well, the selected label of each group.

As we are working in one dimension, we can define the following criteria to determine the grades of membership of the patterns of the linguistic variable values:

1. All the prototypes obtained from the FCM algorithm belong, with all security, to the linguistic variable value correspondent to the group represented by each prototype (for example, 1-group to very-low, 2-group to low, etc).

2. The patterns placed at the left of the $v_1$-prototype belong, with all security, to the first linguistic variable value (for example, very-low), and the patterns placed at the right of the $v_c$-prototype belong, with all security, to the last linguistic variable value (for example, very-high).

3. One pattern can have, at most, two grades of membership to the linguistic variable values different from zero. For example, if the $x$-pattern is placed between the $v_1$ and the $v_2$ prototypes, the grades of membership of the $x$-pattern to the very-low and the low linguistic variable values will be larger than zero, but the grades of membership to the others linguistic variable values will be zero.

Mathematically, the first and the second criteria can directly be expressed. Let us see how we can express the third criterion: Figure 5 represents a hypothetical membership function obtained with the FCM algorithm respect of a hypothetical $v_j$-prototype ($v_j = 10$ in the figure). Let us assume $v_{j+1} = 20$ in the figure. Let us take an element of the sample ($x$) placed on the axe of abscises and the line separating the grey painted area from the dark painted area.

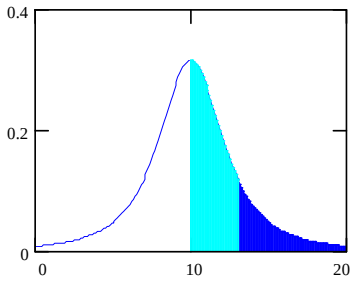


Figure 5

Let us define the following magnitude:

$$w_{xj} = \frac{dark\ area}{grey\ area + dark\ area}$$

When $x \to v_{j+1}$, $w_{xj} \to 0$. When $x \to v_j$, $w_{xj} \to 1$. Moreover, using a geometrically symmetrical criterion for the patterns placed at the left of the prototypes, and defining $w_{xj} = 0 \quad if \quad \left(x < v_{j-1}\ or\ x > v_{j+1}\right)$, the magnitude $w_{xj}$ satisfies the second and the third criteria to determine the membership functions of the values of the linguistic variables, but taking only into account one group. We need calculate the previous magnitude as relative to determine the membership functions searched for.

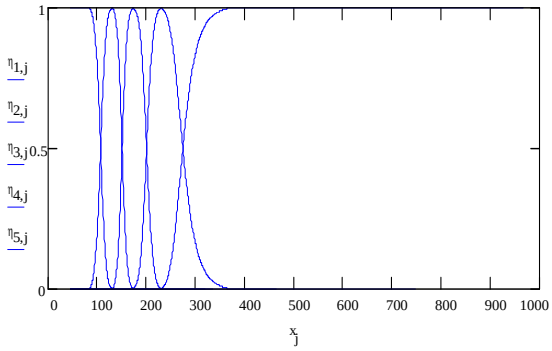


Figure 6. Membership functions of the linguistic variable values corresponding to figure 1.

The $\mu$-membership functions using one-dimensional samples are expressed as:

$$\mu_{xk} = \frac{\sqrt{\sigma_k^{-1}}}{\left(1 + \sigma_k^{-1}\left(x - v_k\right)^2\right)^{3/2}}$$

$\sigma_k$ being the $k$-group fuzzy variance. The areas under the $\mu$-membership functions are very easy to calculate because:

$$\int \frac{\sqrt{b}}{\left(1 + b(z-a)^2\right)^{3/2}} = \sqrt{b}\,\frac{z-a}{\sqrt{1 + b(z-a)^2}} + cte$$

Let us define the following function:

$$\mathrm{A}(z;a,b) = \frac{z-a}{\sqrt{1 + b(z-a)^2}}$$

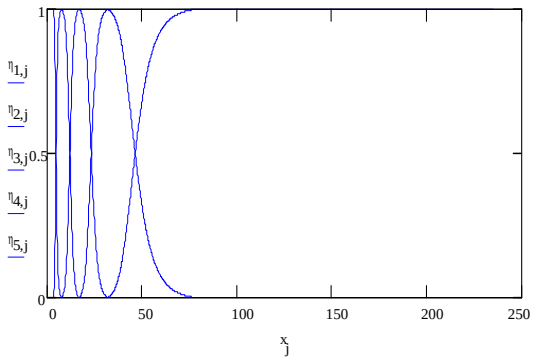


Figure 7. Membership functions of the linguistic variable values corresponding to figure 2.

The $w$-magnitudes are calculated as follows:

$$if\ x < v_1 \Rightarrow w_{x1} = 1, w_{xj} = 0, 1 < j \le c$$

$$if\ x > v_c \Rightarrow w_{xc} = 1, w_{xj} = 0, 1 \le j < c$$

$$if\ v_1 \le x \le v_c \Rightarrow (for\quad 1 \le j \le c)$$

$$\Rightarrow \left\{ \begin{array}{l} if\ v_{j-1} \le x \le v_j \Rightarrow w_{xj} = \dfrac{A\left(x; v_j, \sigma_j^{-1}\right) - A\left(v_{j-1}; v_j, \sigma_j^{-1}\right)}{A\left(v_j; v_j, \sigma_j^{-1}\right) - A\left(v_{j-1}; v_j, \sigma_j^{-1}\right)} \\ if\ v_j \le x \le v_{j+1} \Rightarrow w_{xj} = \dfrac{A\left(v_{j+1}; v_j, \sigma_j^{-1}\right) - A\left(x; v_j, \sigma_j^{-1}\right)}{A\left(v_{j+1}; v_j, \sigma_j^{-1}\right) - A\left(v_j; v_j, \sigma_j^{-1}\right)} \\ if\ x < v_{j-1}\ and\ x > v_{j+1} \Rightarrow w_{xj} = 0 \end{array} \right\}$$

The $w_{xj}$ are defined in such a way that they could be considered as absolutes membership probabilities to the linguistic variable values. To calculate the relative probabilities, we are going to use the Bayes theorem. The probability densities will be $w_{xk} \Big/ \sum_{z \in X} w_{zk}$. We need know the a priori probabilities to apply the Bayes theorem. In [4] and [5], we have given to the membership functions calculated using the FCM algorithm the meaning of densities. Figures 3 and 4 are with this interpretation. The largest density of each group is at each prototype, and it decreases when we go away from the prototype. This is because the prototype represents the group. Increasing the number of groups, the densities at the prototypes will be on a curve representing the sample density. The "mass" belonging to each group will be $\sum_{x \in X} \mu_{xk}$. We take the a priori probabilities as proportional to $\sum_{x \in X} \mu_{xk}$. In consequence, the a posteriori probabilities will be:

$$\eta_{xk} = \frac{w_{xk}}{\sum_{z \in X} w_{zk}} \sum_{y \in X} \mu_{yk} \Bigg/ \sum_{j=1}^{c} \frac{w_{xj}}{\sum_{z \in X} w_{zj}} \sum_{y \in X} \mu_{yj}$$

The a posteriori probabilities are the membership functions searched for. Notice that a probability is a grade of membership, but vice versa is not correct. Figures 6 and 7 represent the membership functions of the linguistic variable values corresponding to figures 3 and 4.

Using similar criterions, the membership functions of the linguistic variable values can be determined when we use another classification technique.

## 3. CONCLUSIONS.

We have obtained the fuzzy sets defining the linguistic variable values from one-dimensional samples using fuzzy clustering techniques. As a consequence, the fuzzy sets are adjusted to the sample distributions. The results of this contribution can be used to apply the fuzzy inference techniques as in the financial sector as another sector. This implies the previous application of the knowledge acquisition techniques. We have applied the results of this contribution to the sale possibility assessment of financial products, which will be the object of another contribution.

## 4. REFERENCES.

[1] Duda P., Hart P., Pattern Classifications and Scene Analisys, Wiley, New York, 1973.

[2] Egea Krauel C, Análisis Estratégico del Sector de Cajas de Ahorro en España, Tesis Doctoral, Universidad Autónoma de Madrid, 1988.

[3] Flores-Sintas A.,Cadenas J.M.,Martin F., A Local Geometrical Properties Application to Fuzzy Clustering, Fuzzy Sets and Systems, 100 (1998) 237-248.

[4] Flores-Sintas A.,Cadenas J.M.,Martin F., Membership Functions in the Fuzzy C-Means Algorithm, Fuzzy Sets and Systems, 101 (1999) 49-58.

[5] Flores-Sintas A.,Cadenas J.M.,Martin F., Partition Validity and Defuzzification, Fuzzy Sets and Systems, to be published.

[6] Flores-Sintas A., Cánovas J., García J., Aproximación a la determinación de perfiles de clientes en entidades financieras utilizando técnicas fuzzy, Actas del VIII Congreso Español sobre Tecnologías y Lógica Fuzzy, (Pamplona, Spain, 1998) 431-435.

[7] Zadeh L., Fuzzy Sets, Inform. and Control, 8-3 (1965) 338-353.