A Multi-level Thresholding-Based Method to Learn Fuzzy Membership Functions from Data Warehouse

Dario Rojas[1], Carolina Zambrano[1], Marcela Varas[2], and Angelica Urrutia[3]

[1] Depto. de Ingeniería Informática y Ciencias de la Computación, Universidad de Atacama, Copiapó, Chile
{dario.rojas,carolina.zambrano}@uda.cl
[2] Depto. de Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción, Concepción, Chile
mvaras@udec.cl
[3] Depto. de Computación e Informática. Universidad Católica del Maule, Talca, Chile
aurrutia@spock.ucm.cl

Abstract. Learn fuzzy membership functions automatically for characterization and operation of fuzzy measures in Data Warehouse is a problem of recent concern. This paper presents a new method to learn membership functions of linguistic labels of fuzzy measures from Data Warehouse. We proposed a multilevel thresholding based method with clustering validation indices in order to obtain optimal number of labels and parameters of membership functions. Validation is performed by comparing the proposal against a supervised learning approach based on clustering and genetic algorithms, including the application in response to queries in a Data Warehouse with fuzzy measures.

Keywords: Fuzzy Logic, Data Warehouse, Multi-Level Thresholding, Clustering, Clustering Validation Indices.

1 Introduction

Most events are vague or uncertain, ie, they imply on its characteristics a certain degree of imprecision (*fuzzyness*). This imprecision may be associated with any type of data as shape, position, time, color, texture, or even the semantics to describe what they are. In many cases, the same concept can have different meanings in different contexts or moments. A warm day in winter is not exactly the same as a warm day in spring, the exact definition of when the temperature goes from warm to temperate is imprecise and context-dependent. It is difficult to associate a specific and unique value with warm or temperate, it can be 24°C, but 25°C could be warm too. This kind of imprecision or *fuzzyness* is constantly linked to phenomena, and is common in every field of study: sociology, physics, biology, finance, engineering, and so on.

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 664–674, 2011.

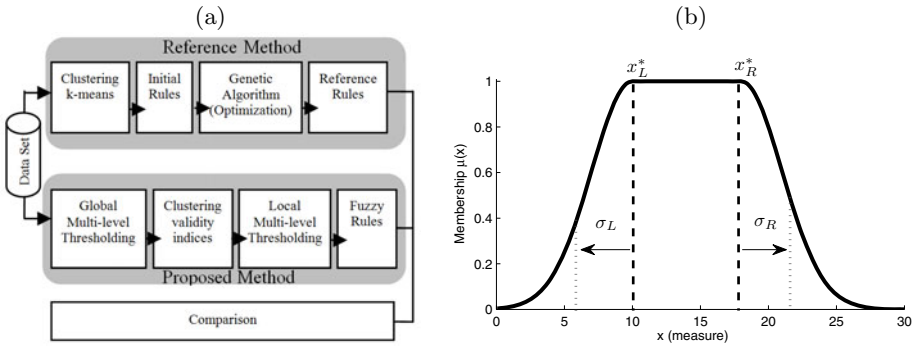

Fig. 1. (a) General scheme for definition and validation of rules generated through proposed method, (b) Example of two-sided Gaussian function

A formalization to express and operates this kind of data, is fuzzy set theory, introduced by L. A. Zadeh [3] in 1965. His proposal considers that each element has a degree of belonging to a set, and this degree is usually a value from 1 (completely belonging) and 0 (not belonging).

Majority of data (precise and imprecise) actually is stored in transactional databases. In order to manage the uncertainty in transactional data bases there are proposals on Fuzzy Databases (FDB), which aim to apply the theory of fuzzy sets to the database, usually as an extension to relational database technology. FDB has been studied at modeling level in [4,13], and in term of design and implementation in [4,9,10,11,12]. One factor to note about the fuzzy management in databases is that they have allowed the management of qualitative information. On the other hand, Data Warehouse (DW) is a repository of data from different sources and usually these sources are transactional databases that collect information over time. DW processes this information and uses it to perform data analysis at the strategic and support decision-making levels in an organization [8].

Fuzzy Data Warehouse (FDW) is defined for purposes of this research as: *A DW that can store data and operate fuzzy measures of a cube*. In addition, one of the main characteristics of FDW is that it can provide qualitative information through fuzzy measures enriched with linguistic labels that are assigned to each indicator according to its value and set of membership function based on the principles of fuzzy logic.

On the other hand, fuzzy multidimensional models, syntax and semantics for answering fuzzy queries have been proposed [1,2]. In this context, an area that has been little explored is the development of rules that explain the nature of data, ie, to obtain the parameters of membership functions from data analysis. This implies that a membership function which defines a linguistic label of an attribute, gets its parameters from context, given from a historical set of data. For the above techniques you can use machine learning and pattern recognition in general. In [16] clustering algorithms and optimization techniques such as hill-climbing is performed in order to obtain classification rules of fuzzy logic,

however, this method is a supervised learning approach, which implies the use of a set of training data to obtain the membership functions, and this is not possible directly in the context of Data Warehouse. On the other hand, in [4] we can find an approach that uses fuzzy clustering to obtain the rules, but this approach only generates triangular membership rules that can be locally optimal [10], because it defines the degree of membership of a cluster based on the distance to its center. In this context, in [14,15] multi-level thresholding techniques are used in order to perform a segmentation of irregular histograms over biofilm images, where the algorithms used are efficient at runtime and optimal, a nice feature, allowing direct and objective comparison of results.

This article proposes a new method in order to obtain membership functions to perform labeling of fuzzy measures in a Data Warehouse using an optimal and efficient unsupervised learning approach, which is organized as follow: In section 2 we present methodology to obtain and validate the new method for obtaining fuzzy rules. In section 3 the application of the proposed method for automatic labeling of fuzzy measures in a Data Warehouse is shown, presenting the results of two common fuzzy queries in such systems. Finally, in section 4 we present the conclusion, comments and future works.

2 Methodology

The proposed method mainly consists in the application of multi-level thresholding and clustering validity indices algorithms in order to obtain the amount and parameters of fuzzy membership functions (two-sided Gaussian functions). In our proposed approach, member functions that are obtained are results of an automatic, unsupervised, optimal and efficient process, and therefore no-subjective and comparable.

In order to validate proposed method, a reference method is developed based on clustering techniques and genetic algorithms (supervised learning approach) in order to ensure high precision of classification through membership functions (also called rules in classification process). Then, results are compared between the proposed method and reference method. In Fig. 1(a), the general scheme for definition and validation of proposed method is depicted.

2.1 Data Set

In order to validate rules obtained from our proposed method, a classical benchmark problem in pattern classification (Fisher's Iris Data Set) is used [5]. The iris data set consists of a set of 150 data samples that map four input features values (sepal length, petal width, petal length y sepal width) into one of three species of iris flowers: Iris-setosa, Iris-versicolor, Iris-virginica.

A DW measure is a quantitative attribute which is mapped in a multidimensional space through dimension hierarchies (qualitative attributes). However, a DW measure is a one-dimensional attribute under pattern recognition approach. Therefore, in order to obtain fuzzy rules and compare results between proposed

method and reference method, a Principal Component Analysis (PCA) is performed in order to project the four-dimensional space into a one-dimensional space and transform iris data set into a DW measure.

2.2 Reference Method

In order to obtain initial fuzzy rules, k-means clustering algorithm is applied in reduced feature space. Fig. 2(a) shows cluster results where symbol • represents Iris-Setosa class, symbol $*$ represents Iris Virginica and symbol $+$ represents Iris Versicolor. For each cluster c_i the mean m_i and standard deviations σ_i are obtained, where $i \in [1, k]$, and $k = 3$ for each flower class. For each cluster, a rule of classification is generated, where each m_i represents the values of data set which have associated a degree of membership equals to 1. On the other hand, each standard deviation are used as right and left parameters for the two-sided Gaussian functions. In Fig. 1(b) is depicted a typical Gaussian functions and the four parameters used: two standard deviations and two mean.

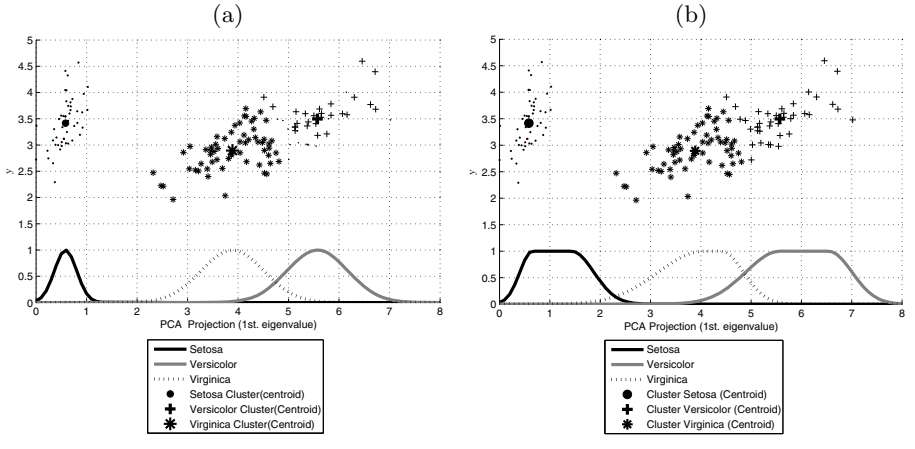

Fig. 2. (a) Two-sided Gaussian Membership Functions obtained through k-means clustering without optimization, (b) Two-sided Gaussian membership functions optimized through genetic algorithm (projection in y is only for visualization purposes)

Initially, each rule is generated from each cluster c_i with $m_i = x^*_{L_i} = x^*_{R_i}$ and $\sigma_i = \sigma_{L_i} = \sigma_{R_i}$. For reference method, three initials rules are generated trough this method, where each rule defines the degree of the values projected by PCA to each class of flowers. The initial rules without optimization are depicted in Fig. 2(a), which have a 79% of precision in classification.

In a second process, a genetic algorithm optimization process is performed in order to optimize initials m_i and σ_i parameters. The chromosomes are a feature vector of four parameters for the three rules (12 chromosomes in total). The fitness function to minimize is normalized error of classification $error =$

$1 - precision$. The rules obtained after optimization process is depicted in Fig. 2(b), where precision of classification is 98%.

2.3 Proposed Method

After of dimensional reduction process (PCA), a relative frequency histogram h is obtained. Then, three multi-level threshold algorithms [18] are applied to h: Entropy-Based thresholding (ENTROPY), Otsu's thresholding (OTSU) and Minimun Error thresholding (MINERROR). In this context, a multi-level thresholding process with $k-1$ thresholds, have a direct relationship with the number of classes k, in which a histogram is partitioned [14]. Viewing thresholding as a problem of clustering frequency histogram h, clustering validity indices [17] can be used in order to obtain the best number of classes k in which the histogram can be clustered, and hence the best number of membership functions or labels can be obtained. In this work, four clustering validity indices are used to determine the best number of thresholds and select the best thresholding technique: Davies-Bouldin Index (DB), Dunn's Index (DN), Index I (IndexI), Calinski Harabasz Index (CH), Xie-Beni Index (XB).

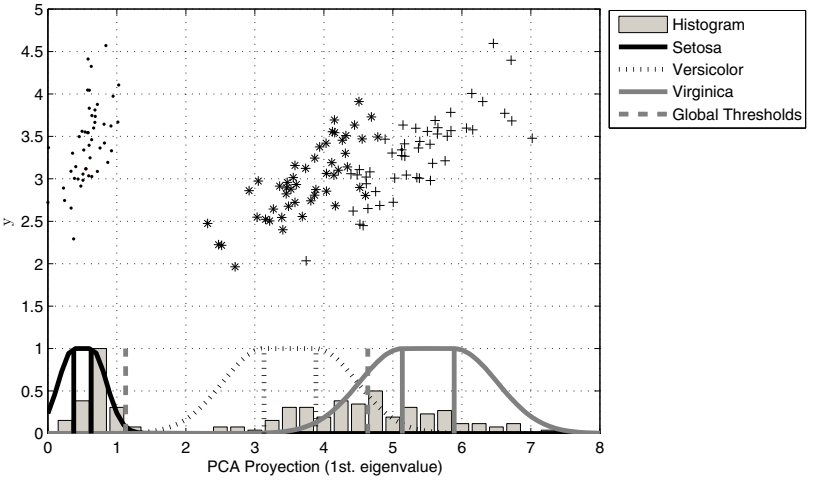

Fig. 3. Two-sided Gaussian membership functions defined through multi-level thresholding techniques (projection in y is only for visualization purposes)

The Fig. 3 shows the resulting rules of multi-level thresholding process. The dotted gray lines represent the global thresholds T_g^G with $g \in [1, k-1]$, which divide the histogram into k initial clusters c_i with $i \in [1, k]$. The solid black lines, dotted black lines and solid gray lines represent the local thresholds T_l^L with $l \in [1, 2]$ for each cluster c_i. Each membership function μ_i for each cluster is defined by:

$$\mu_i = f(T_1^L, T_2^L, \sigma_i), \tag{1}$$

where, f is a two-sided Gaussians function (see Fig. 1(b)), $T_1^L = x_{L_i}^*$, $T_2^L = x_{R_i}^*$ and $\sigma_i = \sigma_{L_i} = \sigma_{R_i}$. As can be see the interval between T_1^L and T_2^L always have a membership degree equal to 1, and the values belonging to the right and left intervals have lower degrees of membership according to the variance of each cluster σ_i.

The classification process carried out with this approach achieves the best accuracy of classification with the MINERROR and OTSU criteria, using an unsupervised technique and obtained objective results, since for the same data always get the same results in an optimal way. Table 1 shows the results of classification precision of the three thresholding criteria using $k = 3$.

Table 1. Precision for each multilevel thresholding criteria, best values are showed in boldface

Thresholding Criteria	ENTROPY	MINERROR	OTSU
Precision	0.5667	**0.9333**	**0.9333**

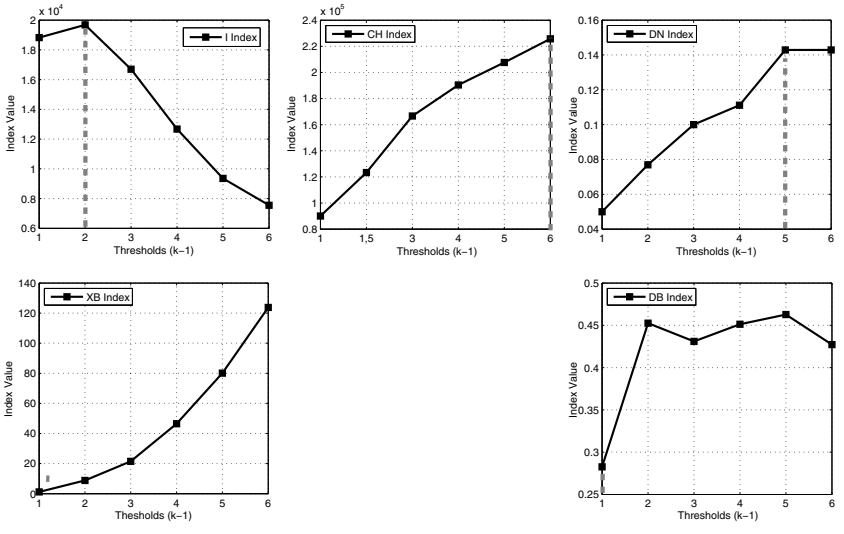

Fig. 4. I, CH, DN, XB and DB cluster validity indices applied after OTSU thresholding. The gray dotted line show the optimal number of clusters obtained by each index.

According to the behavior of cluster validation indices shown in Fig. 4 and Table 2, we can see that CH, XB and DN indices increase monotonically as it increases the numbers of $k - 1$ thresholds used. From the above, the selection criteria that obtain the numbers of thresholds and hence also the number of labels, can be defined by the next expression:

$$k = min(I_j^*, DB_j^*) + 1, \tag{2}$$

where $i \in [1, K]$, I_j^* is the number of thresholds defined by I index, DB_j^* is the optimal number of thresholds obtained by DB index, and K is maximum number of thresholds to find in data set (parameter defined by user).

In summary, the proposed method involves the application of multi-level thresholding algorithm (OTSU) to obtain k membership functions, with k given by Equation 2 through applying clustering validation indices, each membership function μ_i, is obtained by the definition given in Equation 1 through applying global and local thresholding criteria in order to perform the parameterization of a two-sided Gaussian distribution function. We should note that the proposed method is not comparable to other classification techniques directly, because the process itself is for the membership functions and the classification is only for validation purposes, because is not possible to determine classes on measures of a Data Warehouse directly.

Table 2. I, CH, DN, XB and DB cluster validity indices applied with OTSU thresholding. Values shown in boldface represent optimal number of clusters for each index.

# Thresholds	I	CH	DB	DN	XB
1	18822	90066	**0.283**	0.050	**1.10**
2	**19682**	123331	0.452	0.077	8.70
3	16695	166604	0.431	0.100	21.4
4	12672	190410	0.451	0.111	46.4
5	9355	207642	0.463	**0.143**	80.0
6	7549	**225759**	0.427	0.143	123.7

3 Fuzzy Queries in Data Warehouse

In order to perform fuzzy queries using the automatic fuzzy functions proposed, cube depicted in Figure 5 was developed over a subset from a data warehouse system implemented for research at the University of Atacama, Chile [7]. This figure shows part of a conceptual scheme of a DW with fuzzy measures [4]. This schema has been modeled through an instance of the Fuzzy CWM OLAP Meta Model [2]. The cube has six fuzzy measures: marks, age, mathematics, history, science and language, where the last four measures are score of student in a set of tests performed in the admission process of student to the university. In the same context, the dimensions of analysis are: Courses, Time, Students, Cohort and Undergraduate Program.

3.1 Case Study: Automatics Fuzzy Rules from Data Warehouse

Fuzzy rules for the six fuzzy measures obtained through method proposed are depicted in Fig. 6. According the conceptual model, labels are defined according

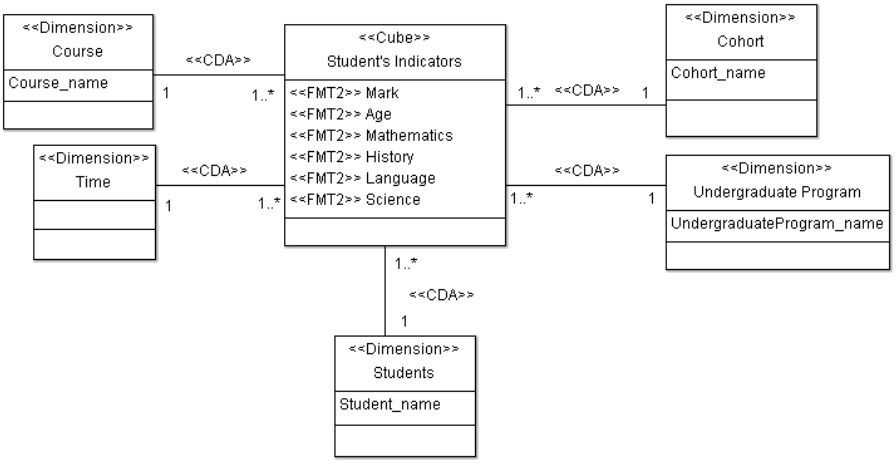

Fig. 5. Conceptual scheme of DW with fuzzy measures. The label $<< FMT2 >>$ is stereotype of Fuzzy Measure (type 2) defined in [2].

the number of rules determined for each measure as: low, medium-low, medium-high, and high. In order to show the use of rules obtained through this method a set of typical queries are performed in the DW using fuzzy approach proposed in [6]:

Query 1 - Average High Marks by Cohort: Table 3 presents the results of query 1, where you can see that all cohorts with high marks have the maximum possibility around 5.9. However, we can appreciate that other values are possible, ie, the 2010 cohort has an average of 5.43 with possibility 84.9%.

Table 3. Results of Query 1: *Average High Marks by Cohort*

Possibility	2003	2004	2005	2006	2007	2008	2009	2010
84.9%	5.69	5.54	5.48	5.52	5.54	5.56	5.49	5.43
93.0%	5.83	5.93	5.80	5.75	5.67	5.65	5.59	5.50
98.2%	5.76	5.67	5.79	5.76	5.76	5.55	5.80	5.73
100%	5.91	5.90	5.90	5.89	5.89	5.87	5.92	5.85

Query 2 - Average High Age by Undergraduate Program: Table 4 presents the results of query 2, where you can see that all undergraduate programs have 24 years as the most possible high age, except Geological Engineering and Business that have values slightly lower.

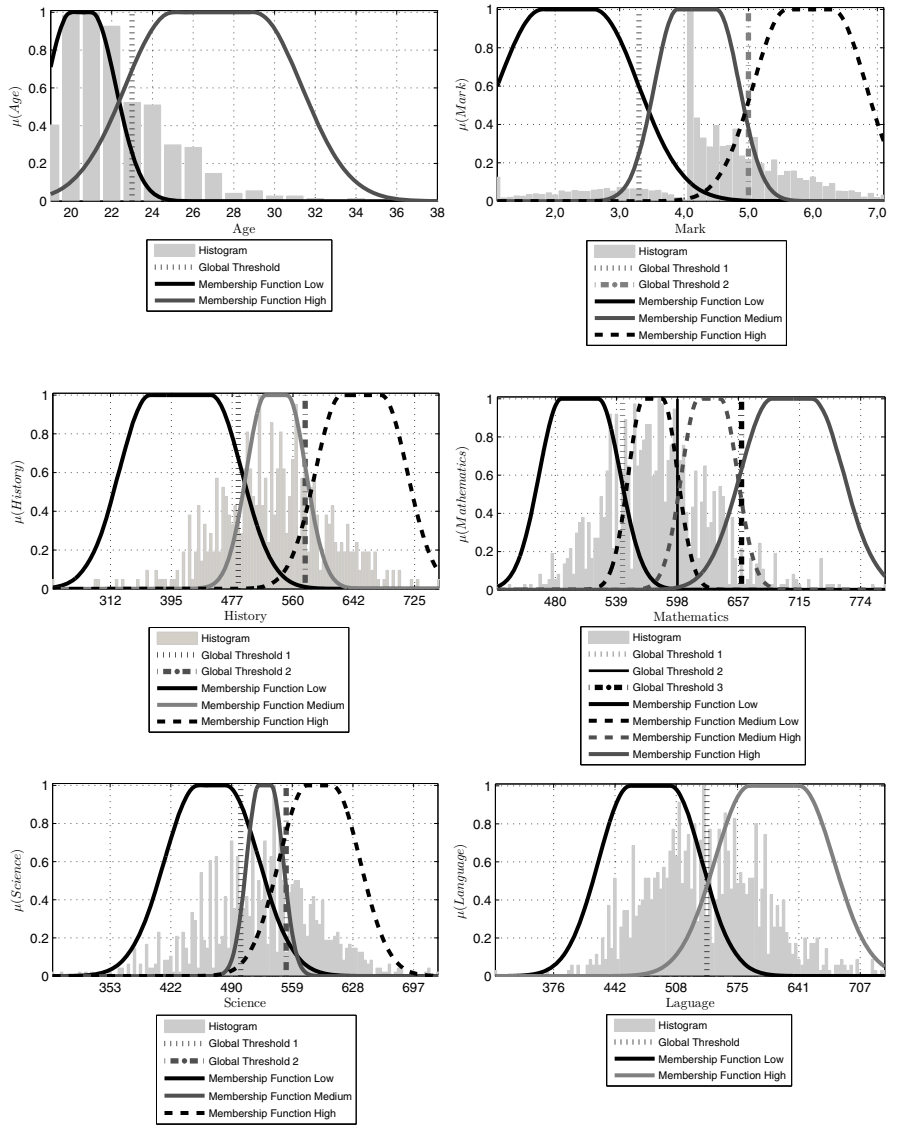

Fig. 6. Fuzzy membership functions DW obtained through multi-level thresholding based method from six measures

Table 4. Results of query *Average High Age by Undergraduate Program.*

Possibility	Business	Computere Science	Geological Engineering	Industrial Engineering	Metallurgical Engineering	Mining Engineering
71.0%	21.0	24.3	21.5	24.7	24.5	22.5
91.8%	22.0	24.3	22.2	24.3	24.5	24.2
100%	23.5	24.0	23.3	24.2	24.0	24.1

4 Conclusion

This work has presented a multi-level thresholding method for obtaining membership functions from fuzzy measures in a Data Warehouse. The proposal has been validated by external criteria (classification rules) and relative criteria (Index Validation) and applied to a real database.

From the results, you can see the great potential of this approach. Undoubtedly, the automation of obtaining the membership functions is one of the least covered issues in fuzzy logic and data warehouse, since it is a process that has not been fully automated.

It is important to add that defining membership functions in fuzzy measures regardless of the context of data, could lead to obtain membership functions with parameters that do not reflect the reality of the domain or organization. For example, the parameters that define the membership function for the high marks label are not the same at a university or another, but should be directly related to the data from each of the universities.

As future works, this method can be appliied to fuzzy levels and extending method to other types of fuzzy data in DW considering the efficiency and objectivity of the proposed approach.

Acknowledgments. This work has been partially supported by MIDAL, Machine Learning and Data Analysis Laboratory and the University of Atacama (University Grant for Research and Artistic Creativity (Projects: *Data Warehouse Difuso Para Análisis Con Jerarquías Difusas* and *Un nuevo algoritmo óptimo de multilevel thresholding para la segmentación de datos e imágenes con determinacin automática de la cantidad de umbrales en tiempo eficiente*)).

References

1. Delgado, M., Molina, C., Sánchez, D., Vila, A., Rodr'iguez-Ariza, L.: A Fuzzy Multidimensional Model for Supporting Imprecision in OLAP. In: Proceedings of IEEE International Conference on Fuzzy Systems (2004)
2. Carrera, S., Varas, M., Urrutia, A.: Transformación de Esquemas Multidimensionales Difusos desde el Nivel Conceptual al Nivel Lógico. Ingeniare. Revista Chilena de Ingenier'ia 18, 165–175 (2010)
3. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
4. Galindo, J., Urrutia, A., Piatinni, M.: Fuzzy Databases: Modeling, Design and Implementation. Idea Group Inc. (2006)

5. Xu, D.: Clustering. IEEE Press Series on Computational Intelligence. A John Wiley & Sons (2009)
6. Rundensteiner, E., Bic, L.: Aggregates in posibilistic databases. In: Proceeding of the 15th Conference in Very Large Databases (VLDB 1989), Amsterdam, Holland, pp. 287–295 (1989)
7. Zambrano C., Rojas D.: Data Warehouse para analizar el comportamiento académico, In: XXIV Congreso de la Sociedad Chilena de Educación en Ingeniería, SOCHEDI 2010, Valdivia, Chile (2010)
8. Inmon, W.: Building the Data Warehouse. John Wiley & Sons (2002)
9. Galindo, J., Urrutia, A., Carrasco, R., Piattini, M.: Relaxing Constraints in Enhanced Entity-Relationship Models Using Fuzzy Quantifiers. IEEE Transactions on Fuzzy Systems 12, 780–796 (2004)
10. Galindo, J., Urrutia, A., Piatinni, M.: Handbook of Research on Fuzzy Information Processing in Databases, Universidad de Málaga, Spain (2008)
11. Galindo, J., Urrutia, A., Carrasco, R., Piattini, M.: Fuzzy Constraints using the Enhanced Entity-Relationship Model. In: XXI International Conference of the Chilean Computer Science Society, Chile, pp. 86–94 (2001)
12. Galindo, J., Urrutia, A., Carrasco, R., Piattini, M.: Relaxing Constraints in Enhanced Entity-Relationship Models Using Fuzzy Quantifiers. IEEE Transactions on Fuzzy Systems 12, 780–796 (2004)
13. Urrutia, A.: Definición de un Modelo Conceptual para Bases de Datos Difusas. Doctoral Thesis (2003)
14. Rojas, D., Rueda, L., Urrutia, H., Ngom A., Carcamo G.: Automatic Segmentation Methods and Applications to Biofilm Image Analysis. In: Data Mining in Biomedical Signaling, Imaging and Systems. CRC Press (2011)
15. Rojas, D., Rueda, L., Urrutia, H., Ngom, A., Carcamo, G.: Image Segmentation of Biofilm Structures Using Optimal Multi-Level Thresholding. International Journal of Data Mining and Bioinformatics 5, 266–286 (2011)
16. Dubois, D., Prade, H., Yager, R.: Fuzzy Information Engineering. Wiley Computer Publishing (1997)
17. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 1650–1655 (2002)
18. Rueda, L.: An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 602–611. Springer, Heidelberg (2008)