Improving Gene Selection in Microarray Data Analysis using Fuzzy Patterns inside a CBR System

Florentino Fdez-Riverola[1], Fernando Díaz[2], Juan M. Corchado[3], Jesús M. Hernández[4] and Jesús San Miguel[4]

[1] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
riverola@uvigo.es
[2] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es
[3] Dept. de Informática y Automática, University of Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es
[4] Dept. Hematología, Hospital Universitario de Salamanca and Centro de Investigación del Cáncer (CIC), Universisty of Salamanca-CSIC, Campus Miguel de Unamuno, 37007, Salamanca, Spain
jmhr@usal.es

Abstract. In recent years, machine learning and data mining fields have found a successful application area in the field of DNA microarray technology. Gene expression profiles are composed of thousands of genes at the same time, representing complex relationships between them. One of the well-known constraints specifically related to microarray data is the large number of genes in comparison with the small number of available experiments or cases. In this context, the ability of identifying an accurate gene selection strategy is crucial to reduce the generalization error (false positives) of state-of-the-art classification algorithms. This paper presents a reduction algorithm based on the notion of fuzzy gene expression, where similar (co-expressed) genes belonging to different patients are selected in order to construct a supervised prototype-based retrieval model. This technique is employed to implement the retrieval step in our new gene-CBR system. The proposed method is illustrated with the analysis of microarray data belonging to bone marrow cases from 43 adult patients with cancer plus a group of three cases corresponding to healthy persons.

1 Introduction and Motivation

Practically all cells in the human body have the same genes, but these genes can be expressed differently at different times and under different conditions. Studying these various states helps scientists understand more about how the cells function and about what happens when the genes in a cell do not work properly. In the past, scientists have only been able to conduct such genetic analyses on a few genes at once. However, in recent years the DNA microarray technology has become a fundamental tool

in genomic research, making the investigation of global gene expression of all aspects of human disease possible [1-4]. Nowadays, it is possible to monitor simultaneously the expression levels of thousands of genes during important biological processes and across collections of related samples.

Microarray technology is based on a database of over 40,000 fragments of genes called expressed sequence tags (ESTs), which are used to measure target abundance using the scanned intensities of fluorescence from tagged molecules hybridised to ESTs [5, 6]. Since the number of examined genes in an experiment it is in term of thousands, different data mining techniques have been intensively used to analyse and discover knowledge from gene expression data [7, 8]. However, having so many fields relative to so few samples creates a high likelihood of finding false positives. This problem is increased if we consider the potential errors that can be present in microarray data, namely *symmetric* and *random* errors [9]. Symmetric (controllable) errors produce approximately similar variations at microarray experiments and it can be handled through normalization techniques [10]. Random (uncontrollable) errors cause different degrees of variations in microarray experiments by chance [11]. Considering a bidimensional matrix containing data from different microarray experiments (from different patients, different times in the same individual, or different tissue types within an individual), we have to deal with the previous commented *intra-experimental* and *inter-experimental* variations. Other issues related with the pre-processing stage within the microarray life cycle are well illustrated in the work of [12].

For several years we have been working in the identification of techniques to automate the reasoning cycle of case based reasoning (CBR) systems [13,14]. In this paper, we propose a fuzzy codification for the gene expression levels of each sample based on the discretization of real gene expression data into a small number of fuzzy membership functions. The proposed method is able to generalize samples as a whole, diminishing the effect of both inter and intra experimental variations. The developed method can be used for different measure platforms (RT-PCR, Affymetrix GeneChip, Rosetta oligoarrays, etc.) and serves as a pre-processing step before gene selection and clustering methods, as we will see later.

We are interested in the development of a robust case-based reasoning system that may be employed in the study of cancer treatment. The goal of the decision support tool is to facilitate the construction of therapies, including the level of aggressiveness of treatment, to more closely match the underlying disease, hopefully reducing side effects in low risk cases and increasing cure rates in high-risk cases.

Input space reduction is often the key phase in the building of an accurate classifier [15]. Based on the fuzzy discretization method presented in this paper, we propose the use of a fuzzy prototype-based retrieval system able to differentiate several kinds of cancer for microarray data. In this case, the goal is the identification of an expression profile that can be used to classify the cancer in our CBR system.

The paper is organized as follows: Section 2 introduces the use of CBR systems and reviews different gene selection approaches, as well as classification techniques for microarray data analysis. Section 3 explains in detail the proposed fuzzy prototype-based retrieval method. Section 4 discusses the experimental results obtained

with the new gene-CBR system built with the proposed method. Finally, Section 5 gives out the concluding remarks and future work.

2 Related Work

Case-based reasoning is a computational reasoning paradigm that involves the storage and retrieval of past experiences to solve new problems. It is an approach that is particularly relevant in scientific domains, where there is a wealth of data but often a lack of theories or general principles.

The domain of molecular biology can be characterized by substantial amounts of complex data, many unknowns, a lack of complete theories and rapid evolution, where reasoning is often based on experience rather than general knowledge. Experts remember positive experiences for the possible reuse of solutions while negative experiences are used to avoid potentially unsuccessful outcomes. Similar to other scientific domains, problem solving in molecular biology can benefit from systematic knowledge management using techniques from AI. Case-based reasoning is particularly applicable to this problem domain because it (*i*) supports rich and evolvable representation of experiences/problems, solutions and feedback; (*ii*) provides efficient and flexible ways to retrieve these experiences; and (*iii*) applies analogical reasoning to solve new problems [16].

Several methods derived from machine learning have been applied to reduce dimensions in the field of microarray data. These works include the application of genetic algorithms [17], wrapper approaches [18], support vector machines [19], etc. Other approaches focus their attention on redundancy reduction and feature extraction [20, 21], as well as the identification of similar gene classes making prototypes-genes [22]. One way or another, the selected method has to pursue two main goals: (*i*) reduce the cost and complexity of the classifier and (*ii*) improve the accuracy of the model.

Classical reduction dimension methods applied to microarray data [23] tend to identify differentially expressed genes from a set of microarray experiments. A differentially expressed gene is a gene which has the same expression level for all examples of the same class, but different for those examples belonging to different classes. The relevance value of a gene depends on its capacity of being differentially expressed. However, a non-differentially expressed gene will be considered irrelevant and will be removed from the classification process even though it might well contain information that would improve the classification accuracy.

The task addressed here is slightly different from that of feature selection for gene expression based classifiers [24, 25]. Our proposed method aims to find all genes that are significantly expressed between the existing classes in order to obtain a fuzzy representation of the expression levels belonging to those genes that best explain each class in the form of a fuzzy-prototype. The final goal is the application of the proposed method as a retrieval step for our gene-CBR system.

3 Fuzzy Prototype-based Retrieval Method for CBR systems

The proposed method employs a fuzzy codification for the gene expression levels of each case, based on the discretization of real gene expression data into a small number of fuzzy membership functions. The whole algorithm comprises of two main steps. First, we discretize the gene expression levels into binary variables according to a supervised learning process. Then, a fuzzy pattern is generated from the data, which is representative for each specific pathology. To carry out the integration of the proposed method within the CBR life cycle, a measured distance has to be defined in order to determine the distance of a gene expression profile (or new case) to a specific gene expression pattern.

3.1 Fuzzy Discretization of Gene Expression Levels

Given a set of n features or attributes (in this work, gene expression levels), $\boldsymbol{F} = \{F_1, F_2, ..., F_n\}$, the discretization process is based on determining the membership function of each feature to three linguistic labels (LOW, MEDIUM, and HIGH). Then, each real value F_j is replaced by its three values of membership to these fuzzy labels (μ_{jL}, μ_{jM} and μ_{jH}, respectively), and so, a new set of $3n$ features, $\boldsymbol{F'} = \{\mu_{1L}, \mu_{1M}, \mu_{1H}, ..., \mu_{nL}, \mu_{nM}, \mu_{nH}\}$ is constructed from the original set of features $\boldsymbol{F}$.

The membership functions to linguistic labels are defined in a similar way to the form that has been used by [26, 27]. These authors used a polynomial function that approximates a Gaussian membership function, where its centre and amplitude depend on the mean and on the variability of the available data respectively. The original membership functions are considered symmetric, but, in this work we have considered asymmetric functions for the linguistic labels in the extremes (labels LOW and HIGH). To support this choice, it is assumed that values below the centre of membership function for label LOW are *low* values for the feature F_j at a fuzzy degree of 1. The same consideration is made to the label HIGH.

Concretely, the membership function for the label LOW is defined by:

$$\mu_{jL}(x)=\begin{cases} 1 & \text{if } x-c_{jL}\le 0 \\ 1-2\left(\dfrac{x-c_{jL}}{\lambda_{jL}}\right)^2 & \text{if } 0\le x-c_{jL}\le \dfrac{\lambda_{jL}}{2} \\ 2\left(1-\dfrac{x-c_{jL}}{\lambda_{jL}}\right)^2 & \text{if } \dfrac{\lambda_{jL}}{2}\le x-c_{jL}\le \lambda_{jL} \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

where c_{jL} is the mean of the values of feature F_j below the mean of all values of the feature F_j, (namely, given $c_{jM} = \mathrm{E}[F_j]$, the centre c_{jL} is is the mean of the values of feature F_j that are comprised between $\mathtt{min}(F_j)$ and c_{jM}) and the λ_{jL} parameter is the

distance between c_{jM} and c_{jL}, $\lambda_{jL} = c_{jM} - c_{jL}$. As it is defined, this function is asymmetric, as is shown in Figure 1.

For the label HIGH the definition of its membership function is made in a similar way,

$$\mu_{jH}(x)=\begin{cases}1 & \text{if } x-c_{jH}\geq 0\\ 1-2\left(\dfrac{x-c_{jH}}{\lambda_{jH}}\right)^2 & \text{if } -\dfrac{\lambda_{jH}}{2}\leq x-c_{jH}\leq 0\\ 2\left(1+\dfrac{x-c_{jH}}{\lambda_{jH}}\right)^2 & \text{if } -\lambda_{jH}\leq x-c_{jH}\leq -\dfrac{\lambda_{jH}}{2}\\ 0 & \text{otherwise}\end{cases}. \tag{2}$$

but in this case, the centre c_{jH} is the mean of the values of F_j that are comprised between the mean value of all values, c_{jM}, and the maximum value, $\max\{F_j\}$, whereas the amplitude parameter, λ_{jH}, is given by the difference $c_{jH} - c_{jM}$. This function extends the right side of the domain of Feature F_j, and it is shown in Figure 1. It is also an asymmetric membership function.

Last, the membership function to the label MEDIUM is a symmetric function defined as:

$$\mu_{jM}(x)=\begin{cases}1-2\left(\dfrac{\left\|x-c_{jM}\right\|}{\lambda_M}\right)^2 & \text{if } 0\leq\left\|x-c_{jM}\right\|\leq\dfrac{\lambda_{jM}}{2}\\ 2\left(1-\dfrac{\left\|x-c_{jM}\right\|}{\lambda_{jM}}\right)^2 & \text{if } \dfrac{\lambda_M}{2}\leq\left\|x-c_{jM}\right\|\leq\lambda_{jM}\\ 0 & \text{otherwise}\end{cases}. \tag{3}$$

where the centre parameter, c_{jM}, is the mean of all values of feature F_j, $c_{jM} = \mathrm{E}[F_j]$, and the amplitude parameter λ_{jM} is given by the half of the distance between the centres of the extreme functions, namely, $\lambda_{jM} = ½\ (c_{jH} - c_{jL})$. The form of this function is also shown in Figure 1.

Once defined the three membership functions for each feature F_j, a threshold value Θ can be established (for example, 0.5) to discretize the original data in a binary way, according to any linguistic label from the defined labels LOW, MEDIUM and HIGH. The discriminatory criterion for any label is simply defined by:

$$F'_{j\bullet}=\begin{cases}1 & \text{if } \mu_{j\bullet}(x)\geq\Theta\\ 0 & \text{if } \mu_{j\bullet}(x)<\Theta\end{cases}. \tag{4}$$

As is shown in Figure 1, for concrete values of threshold Θ, specific zones of the feature domain for which none of the labels will be activated can exist (see the neighbour region of the intersection of membership functions of label MEDIUM and HIGH in Figure 1). This fact must be interpreted as the specific value of the feature is not enough to assign it a significant linguistic label at the significance degree of membership fixed by threshold Θ. On the other hand, one value can activate simultaneously two linguistic labels, since at the significance level given by Θ, any assignment of the measure to a linguistic label is significant (see, the neighbour region of the intersection of label MEDIUM and HIGH in Figure 1).

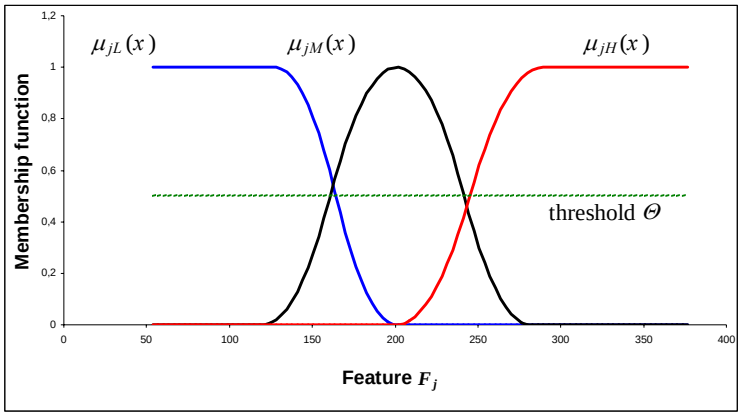

Fig. 1. Membership functions for the linguistic labels: LOW, MEDIUM and HIGH

This section has presented a method used to discretize numeric features into binary variables according to the definition of three linguistic labels, and therefore the method is defined in a fuzzy sets manner. Summarizing, given a data set $\boldsymbol{D}$ with m observations $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ about n numeric features $\boldsymbol{F} = \{F_1, ..., F_n\}$, namely, $\boldsymbol{x}_i \in \mathbf{R}^n$, the fuzzy discretization process, defined above, transforms the original data set into another set with the same number of observations but a different number of features. The new data set $\boldsymbol{D'}$ has m observations which are now referred to as a set of $3n$ binary features, namely, $\boldsymbol{x'}_i \in \{0, 1\}^{3n}$. The real value of feature F_j for the observation $\boldsymbol{x}_i$, denoted by x_{ij}, is replaced by the three binary values given by expression (4) for each linguistic label, that is to say, by the tuple $\langle F'_{jL}(x_{ij}), F'_{jM}(x_{ij}), F'_{jH}(x_{ij})\rangle$.

3.2 Generating Fuzzy Patterns from Data

This section explains how to generate a fuzzy pattern from data, which is representative for a specific decision class. The process is carried out according to a supervised learning process from the available data as described below.

Given a subset of observations $\boldsymbol{D}_i = \{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, ..., \boldsymbol{x}_{i_m}\} \subseteq \boldsymbol{D}$, which have associated the same class label C_i, for any observation $\boldsymbol{x}_{i_l}$ $(i_1 \leq i_l \leq i_m)$,

- First, it is discretized with regard to the linguistic labels LOW, MEDIUM and HIGH associated to each feature, F_j. Namely, the discrete values $F'_{jL}(x_{i_l j})$, $F'_{jM}(x_{i_l j})$, and $F'_{jH}(x_{i_l j})$ are computed using the expression given by (4). Then, the three binary values for each feature are replaced by a single label, $F''_j(\boldsymbol{x}_{i_l}) \in \{\text{L, LM, M, MH, H, *}\}$. If only one of the three binary values is active, the respective label is assigned: L (LOW), M (MEDIUM), and H (HIGH). As mentioned in Section 3.1, a unique real value can activate simultaneously two linguistic labels, so it may occur that two binary values are activated – the possible cases are LM (LOW and MEDIUM) and MH (MEDIUM and HIGH). Finally, it is also possible that one value does not fire any linguistic label, and then, the label * is assigned. The assignment criteria is given completely by expression (5).

$$F''_j(\boldsymbol{x}_{i_l}) = \begin{cases} L & \text{if } F'_{jL}(x_{i_l j}) = 1 \wedge F'_{jM}(x_{i_l j}) = 0 \wedge F'_{jH}(x_{i_l j}) = 0 \\ LM & \text{if } F'_{jL}(x_{i_l j}) = 1 \wedge F'_{jM}(x_{i_l j}) = 1 \wedge F'_{jH}(x_{i_l j}) = 0 \\ M & \text{if } F'_{jL}(x_{i_l j}) = 0 \wedge F'_{jM}(x_{i_l j}) = 1 \wedge F'_{jH}(x_{i_l j}) = 0 \\ MH & \text{if } F'_{jL}(x_{i_l j}) = 0 \wedge F'_{jM}(x_{i_l j}) = 1 \wedge F'_{jH}(x_{i_l j}) = 1 \\ H & \text{if } F'_{jL}(x_{i_l j}) = 0 \wedge F'_{jM}(x_{i_l j}) = 0 \wedge F'_{jH}(x_{i_l j}) = 1 \\ * & \text{if } F'_{jL}(x_{i_l j}) = 0 \wedge F'_{jM}(x_{i_l j}) = 0 \wedge F'_{jH}(x_{i_l j}) = 0 \end{cases}. \quad (5)$$

- Secondly, the fuzzy pattern (corresponding to the class C_i) is constructed from the discretized and summarized data, selecting those labels of features which are different to the label "*" and have an appearance relative frequency in set $\boldsymbol{D}_i$ equal to or greater than a predefined ratio Π ($0 < \Pi \leq 1$, for example, $\Pi = 2/3$). Formally, for each feature F_j, the appearance frequency of any label $E \in \boldsymbol{E} = \{\text{L, LM, M, MH, H, *}\}$ in the set $\boldsymbol{D}_i$, $\pi_{ij}(E)$, can be computed according to the expression given by

$$\pi_{ij}(E) = \frac{\sum_{i_1 \leq i_l \leq i_m} \delta_j(\boldsymbol{x}_{i_l}, E)}{i_m}, \text{ where } \delta_j(\boldsymbol{x}_{i_l}, E) = \begin{cases} 1 & \text{if } F''_j(\boldsymbol{x}_{i_l}) = E \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

Once, the frequency of each label is computed for every feature, a 3-tuple of the form ⟨feature, label, frequency⟩ is included in the fuzzy pattern of class C_i, only if its frequency exceeds the predefined ratio Π. Namely, the fuzzy pattern $\boldsymbol{P}_i$ is given by:

$$\boldsymbol{P}_i = \left\{ \bigwedge_{F''_j \in \boldsymbol{F}''} \left\langle F''_j, E^j, \pi^j \right\rangle : E^j = \arg\max_{E \in \boldsymbol{E}} \{\pi_{ij}(E)\} \wedge E^j \neq * \wedge \pi^j = \pi_{ij}(E^j) \geq \Pi \right\}. \quad (7)$$

The predefined ratio Π controls the degree of exigency for selecting a feature as a member of the pattern, since the higher the value of Π, the fewer number of features which make up the pattern.

The method presented in this section aims to construct a fuzzy pattern which is representative of a collection of observations belonging to the same decision class, namely, the gene expression pattern of a specific kind of cancer. The pattern's quality

of fuzziness is given by the fact that the labels, which make it up, come from the linguistic labels defined during the discretization stage. On the other hand, if a specific label of one feature is very common in all the examples (belonging to the same class), this feature is selected to be included in the pattern and, therefore, a frequency-based criteria is used for selecting a feature as part of the pattern.

3.3 Measuring the Distance of a New Case to a Gene Expression Pattern

This section describes how to measure the distance of a gene expression profile to a specific gene expression pattern. This feature is very important to perform different tasks such as clustering, supervised classification, the recovery of similar cases in a CBR system, and so on.

The defined metric is based on the comparison of the similarity of any two of the linguistic labels defined in Section 3.1. It is assumed that the similarity of two linguistic labels is determined by the degree of overlapping between labels and its definition is argued below.

From the traditional theory of sets it is known that the similarity between two sets A and B (and assuming that set A acts as a reference set), can be evaluated by:

$$sim(A,B) = \frac{|A \cap B|}{|A|}. \tag{8}$$

Likewise, a similarity metric can be defined between fuzzy sets. In this case, it has been considered that the fuzzy intersection of two fuzzy sets A and B (represented by its membership functions, μ_A and μ_B, respectively) is given by the application of the `min` operator to the two membership functions, namely, $\mu_{A \cap B} = \texttt{min}\ \{\mu_A, \mu_B,\}$. On the other hand, the cardinality operator can be replaced by the integral operator, and then the similarity between two fuzzy sets can be evaluated by:

$$sim(A,B) = \frac{\int \min\{\mu_A(x), \mu_B(x)\}dx}{\int \mu_A(x)dx}. \tag{9}$$

The metric $sim(A, B)$ varies between the values 0 (total dissimilarity) and 1 (total similarity). A graphical interpretation of this similarity measure is given in Figure 2.

In this example, it is shown that the similarity of label B with regard to label A grows as the intersection area increases, and vice-versa. At this point, the analytical calculation of the integrals that appear in expression (9) must be made. After some calculus, facilitated by the fact that the defined membership-functions are polynomial, a closed form for these integrals has been determined. These calculations are out of the scope of this work, and they do not contribute to the explanation of our proposal.

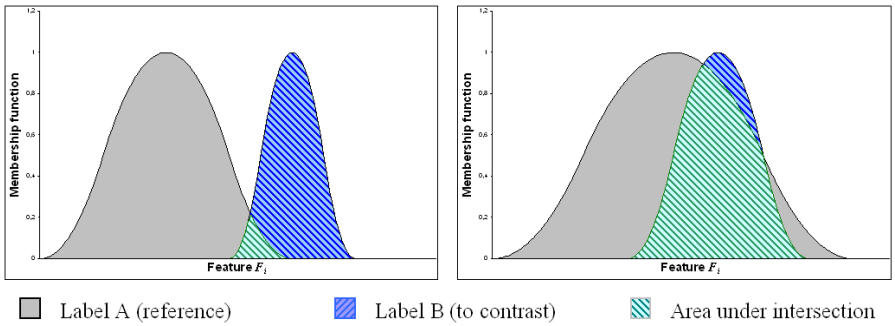

Fig. 2. Relation between the area below the membership function and the similitude of linguistic labels

Now, given a set of data $\boldsymbol{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$, where $\boldsymbol{x}_i \in \mathbf{R}^n$, is a vector of n real values, each one referred to a feature in the set of features provided by a patient's gene expression profile, $\boldsymbol{F} = \{F_1, F_2, ..., F_n\}$. A representative pattern of the data set $\boldsymbol{D}$ can be extracted according to the process described in the previous section, which is an expression of the form:

$$\boldsymbol{P} = \bigwedge_{F''_j \in \boldsymbol{F}''} \left\langle F''_j, E^j, \pi^j \right\rangle = \left\langle F''_{j_1}, E^{j_1}, \pi^{j_1} \right\rangle \wedge \ldots \wedge \left\langle F''_{j_n}, E^{j_n}, \pi^{j_n} \right\rangle . \tag{10}$$

where j_n is the number of variables which the pattern has. Given a new observation $\boldsymbol{x} \in \mathbf{R}^n$, we are interested in evaluating the distance between the observation $\boldsymbol{x}$ and the pattern given by $\boldsymbol{P}$. After discretizing and summarizing the novel observation $\boldsymbol{x}$ following the process described in Section 3.1, the original vector $\boldsymbol{x} = \langle F_1(\boldsymbol{x}), ..., F_n(\boldsymbol{x}) \rangle$ will be replaced by its discrete version, $\boldsymbol{x}'' = \langle F''_1(\boldsymbol{x}), ..., F''_n(\boldsymbol{x}) \rangle$, where $F''_j(\boldsymbol{x})$ is defined by (6).

Then, assuming that the metric given by $sim(A, B)$ is available, the distance between the novel observation $\boldsymbol{x}$ and the pattern $\boldsymbol{P}$, denoted by $d(\boldsymbol{P}, \boldsymbol{x})$, is defined as:

$$d(\boldsymbol{P}, \boldsymbol{x}) = \frac{j_n}{\sum_{j_1 \le j_k \le j_n} sim(E^{j_k}, F''_{j_k}(\boldsymbol{x})) \cdot \pi^{j_k}} - 1 . \tag{11}$$

This definition assumes that the similarity of an observation $\boldsymbol{x}$ to a pattern $\boldsymbol{P}$ depends on the sum of the similarity of their individual labels – evaluated by the term $sim(E_j, F''_j(\boldsymbol{x}))$ – and weighted by term π^j – the relative frequency of the pattern's label for the jth feature, E^j, in the original data set $\boldsymbol{D}$. Then, the distance is defined as inversely proportional to this similarity and normalized by the number of terms of the pattern – to allow us to compare the same observation with patterns of different length – and adjusted in such a way, that the range of the defined distance is between 0 (perfect match) to ∞ (complete dissimilarity).

Finally, it may be interesting to have threshold Δ associated to each pattern, so that the distance between an observation (or other pattern) and a reference pattern exceeds this threshold, it must be concluded that the observation is out of the influence area of

the reference pattern. To compute this threshold, we must consider the mean of the distances of every observation $\boldsymbol{x}_i \in \boldsymbol{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ (which were used to construct $\boldsymbol{P}$) to the pattern $\boldsymbol{P}$, and the threshold is defined as the upper bound of the confidence interval of this mean with a significance level of 5%. Then, the threshold for the pattern $\boldsymbol{P}$, Δ_P, is defined as:

$$\Delta_P = E[d(\boldsymbol{P}, \boldsymbol{x}_i)] + \frac{1{,}96}{\sqrt{m-1}} \sqrt{\mathrm{Var}[d(\boldsymbol{P}, \boldsymbol{x}_i)]} \,. \tag{12}$$

and so, it depends on the mean distance of all observations (used to construct it) to the pattern $\boldsymbol{P}$, the variability of these distances and the number of available observations.

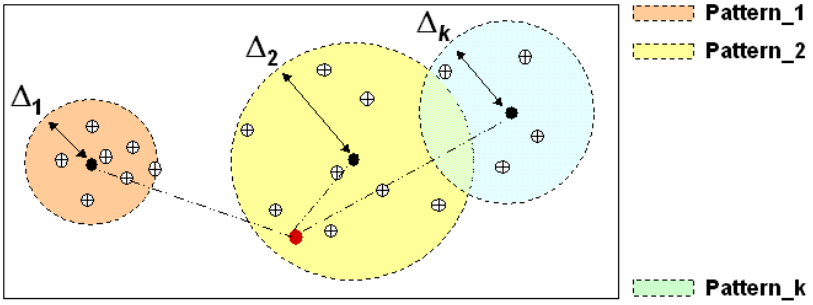

Fig. 3. Graphical interpretation of the threshold Δ

Figure 3 shows a graphical interpretation of the threshold Δ and how a novel observation (the red point) can be classified within the nearest pattern (in the example, it will be assigned to the second pattern).

4 Case Study: Acute Myeloid Leukemia

The study described in this paper was carried out in the context of the FSfRT architecture. FSfRT is a structured hybrid system that can employ several soft computing techniques in order to accomplish the 4-steps of the classical CBR life cycle [28].

The FSfRT architecture is an extension of a previous successful system [29] able to make predictions of red tides (discolourations caused by dense concentrations of microscopic sea plants, known as phytoplankton). The FSfRT architecture allows us the combination of several soft computing techniques in order to test their suitability working together to solve complex problems. The core and the interfaces of FSfRT have been coded in Java language and new capabilities are being developed. The general idea is to have different programmed techniques that are able to work separately and independently in co-operation with the rest. The main goal is to obtain a general structure that could change dynamically depending on the type of problem. Figure 4 shows a schematic view of the system.

The core of the system, which is composed of a *Knowledge Acquisition Module* (KAM), is shown on the left of Figure 4. The KAM is able to store all the information

needed by the different techniques employed in the construction of the final gene-CBR system. In the retrieval and reuse stages, several soft computing techniques can be used [30, 31], while in the revise stage, our system employs a set of TSK fuzzy systems [32] in order to perform the validation of the initial solution proposed by the system.

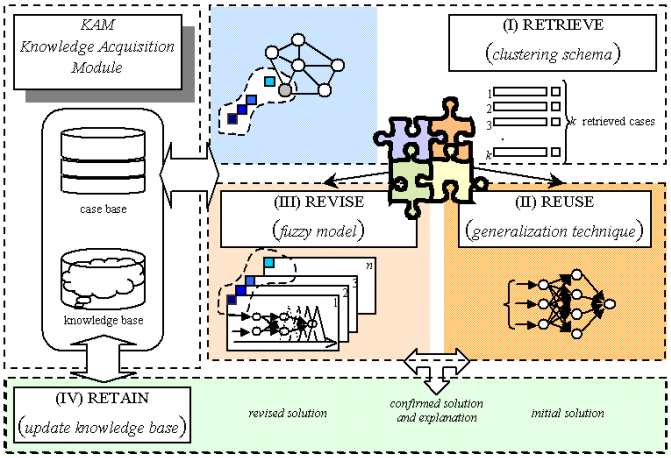

Fig. 4. FSfRT system architecture

The gene-CBR system is being developed, and as a first step, the fuzzy prototype-based retrieval method previously exposed has been evaluated. The main goal is to reduce the original data set of features while maintaining the classification accuracy of the system classifying the cancer.

Recent studies in human cancer have demonstrated that microarrays can be used to develop a new taxonomy of cancer, including major insights into the genesis, progression, prognosis and response to therapy based on gene expression profiles [33]. Often, cancers that appear histologically similar can have dramatically different responses to standard therapies and different courses of development. Since these differences in behaviour are likely to be reflected in the differences in the set of genes expressed, one promising use for microarrays is to more finely differentiate cancers using gene expression levels to bolster standard histology.

In our experiments, we work with a database of bone marrow cases from 43 adult patients with AML, a particular kind of cancer, plus a group of three samples belonging to healthy persons for test purposes (see Table 1 for a concrete description). The group of ill patients can be divided into four different groups, each of them characterized for having a different type of cancer with a different treatment and outcome. Each case (microarray experiment) stores 22,283 ESTs corresponding to the expression level of thousands of genes. The data consisted of 1,025,018 scanned intensities.

Table 1. Classification of patients taking into account the type of cancer

	healthy	APL	AML-inv()	AML-mono	AML-other
Number of patients	3	10	4	7	22

In the group of patients suffering AML-other, it was detected by the experts that new types of cancer would be able to rise. In our experiments, we randomly select 31 cases for training the method and 12 cases for test purposes (38% of the whole data, including at least one example from each group).

In order to generate a fuzzy pattern for each pathology (as described in Section 3.3) without taking into account the healthy people, the first step carried out was to discretize the expression profiles of all the genes regarding the linguistic labels LOW, MEDIUM and HIGH. To do this, several experiments were carried out to select and adequate value for the theta (Θ) threshold (Section 3.1). The next step was to define the minimum appearance frequency, phi (Π), needed to consider a gene for representing a pathology in its corresponding fuzzy pattern (Section 3.2).

Table 2 shows a summary of different values for the theta and phi ratios and the percentage of misclassifications over the test cases.

Table 2. Percentage of misclassifications using the Fuzzy Prototype-based Retrieval Method

	Π = 0.66	Π = 0.75	Π = 0.80	Π = 0.83	Π = 0.86
Θ = 0.75	41.67%	33.33%	33.33%	25.00%	25.00%
Θ = 0.85	41.67%	25.00%	25.00%	8.33%	25.00%
Θ = 0.95	41.67%	16.67%	16.67%	8.33%	16.67%
Θ = 0.975	33.33%	16.67%	16.67%	8.33%	8.33%
Θ = 0.9875	33.33%	16.67%	16.67%	8.33%	16.67%
Θ = 0.99	25.00%	16.67%	16.67%	8.33%	25.00%
Θ = 0.999	16.67%	8.33%	8.33%	**0.00%**	25.00%

Figure 5 also shows a representation of the classification error versus phi and theta values.

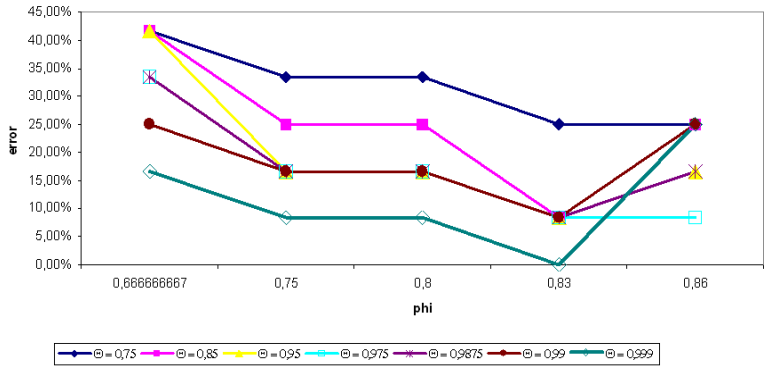

Fig. 5. Classification error varies accordingly phi and theta parameters

From Table 2 and Figure 5 it can be seen that for Θ = 0.999 and Π = 0.83, the proposed method was able to correct classify the whole test bed. Moreover, the proposed method employs on average only 2% of the whole data for this task (see Table 3).

As we mention above, the main goal of our method was to reduce the original set of features while maintaining the classification accuracy of the system classifying the

cancer. In this context, Table 3 shows the gene reduction percentage using the selected phi and theta values. For example, to identify the patients with APL leukemia we only need to analyse 681 variables (genes) out of the 22,283 that compose the whole case (patient descriptor).

Table 3. Reduction percentage obtained over the whole data using optimal values for theta and phi parameters

	APL	AML-inv()	AML-mono	AML-other		
				Sub_1	Sub_2	Sub_3
Original set	22,283	22,283	22,283	22,283	22,283	22,283
Selected set	681	591	292	176	235	817
% reduction	96.9%	97.4%	98.7%	99.2%	98.8%	96.3%

As Table 3 shows, the fuzzy prototype-based retrieval method was able to identify the three subtypes of AML-other as experts previously sensed. In this sense, the outcome generated overcomes those obtained by specific classification techniques such as PAM (*Prediction Analysis of Microarrays*) [34].

The main advantages of the proposed technique are that new subgroups of cancer are correctly identified and that fewer genes are needed in order to classify each case.

These results are very promising considering the reduction percentage of genes done by the proposed technique, especially if this work is compared with the previous one presented in [33]. However, this work that has been developed in the past eight months requires further experimental validation and follow up study. Many current efforts are being directed towards this area of research.

5 Conclusions and Future Work

An advantage of CBR systems as a problem-solving paradigm is that it is applicable to a wide range of problems. It can be used to propose new solutions or evaluate solutions to avoid potential problems. In the work of [35] it is suggested that analogical reasoning is particularly applicable to the biological domain, partly because biological systems are often homologous (rooted in evolution). Also, due to the fact that biologists often use a form of reasoning similar to CBR, where experiments are designed and performed based on the similarity between features of a new system and those of known systems.

In this work, we have presented a fuzzy codification for the gene expression levels of microarray data, based on the discretization of real gene expression data into a small number of fuzzy membership functions. Our proposed method aims to find all genes that are significantly expressed between the existing classes in order to obtain a fuzzy representation of the expression levels belonging to those genes that best explain each class in the form of a fuzzy-prototype. Then the proposed method is able to generalize over all of the samples diminishing drastically the number of genes needed to perform correct classifications. The fuzzy representation technique can be used to implement the retrieval stage of gene-CBR system under construction. Empirical

studies show that this reduction technique allows to obtain a more general knowledge about the problem and to gain a deeper insight into the importance of each gene related to each pathology.

The remaining work is geared towards the study of new techniques that can be used for implementing the reuse, revision and retain phases of the gene-CBR life cycle. It is always important to completely define how a case could be represented and how we can maintain clinical and biological characteristics as well as temporary evolution of all the patients stored in the case base.

References

1. Schena, M., Shalon D., Davis, R., Brown, P.O.: Quantitative monitoring of gene expression patterns with a cDNA microarray. Science, Vol. 270. (1995) 467–470
2. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su Y.A., Trent, J.M.: Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genetics, Vol. 14. (4). (1996) 367–370
3. The Chipping Forecast I. Special Supplement. Nature Genetics, Vol. 21. (1999)
4. The Chipping Forecast II. Special Supplement. Nature Genetics, Vol. 32. (2002)
5. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., Lockhart, D.H.: High density synthetic oligonucleotide arrays. Nature Genetics, Vol. 21. (1999) 20–24
6. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. Vol. 286 (5439). (1999) 531–537
7. Articles on microarray data mining. ACM SIGKDD Explorations Newsletter, Vol. 5 (2). (2003) 1–139
8. Cho, S.B., Won, H.H.: Machine learning in DNA microarray analysis for cancer classification. Proc. of the First Asia-Pacific Bioinformatics Conference, Vol. 19. (2003) 189–198
9. Morrison, N., Hoyle, D.C.: Normalization concepts and methods for normalizing microarray data. In Berrar, D.P., Dubitzky, W., Granzow, M. (eds.). A Practical Approach to MicroArray Data Analysis, Kluwer Academic Publishers, Boston (2003)
10. Bilban, M., Buehler, L.K., Head, S., Desoye, G., Quaranta, V.: Normalizing DNA microarray data. Current Issues in Molecular Biology, Vol. 4 (2). (2000) 57–64
11. Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H.: Normalization strategies for cDNA microarrays. Nucleic Acids Research, Vol. 28 (10). (2000) e47
12. Rubinstein, B.I.P., McAuliffe, F., Cawley, S., Palaniswami, M., Ramamohanarao, K., Speed, T.S.: Machine learning in low-level microarray analysis. ACM SIGKDD Explorations Newsletter, Vol. 5 (2). (2003) 130–139
13. Corchado, J.M., Corchado, E.S., Aiken, J., Fyfe, C., Fdez-Riverola, F., Glez-Bedia, M.: Maximum Likelihood Hebbian Learning Based Retrieval Method for CBR Systems. Proc. of the 5th International Conference on Case-Based Reasoning, (2003) 107–121
14. Corchado, J.M., Aiken, J., Corchado, E., Lefevre, N., Smyth, T.: Quantifying the Ocean's CO2 Budget with a CoHeL-IBR System. Proc. of the 7th European Conference on Case-based Reasoning, (2004) 533–546
15. Cakmakov, D., Bennani, Y.: Feature selection for pattern recognition, Informa Press (2003)
16. Jurisica, I., Glawgow, J.: Applications of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics, Vol. 25 (1). (2004) 85–95

17. Li, L., Darden, T.A., Weinberg, C.R., Levine, A.J., Pedersen, L.G.: Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Combinatorial Chemistry and High Throughput Screening, Vol. 4 (8). (2001) 727–739
18. Blanco, R., Larrañaga, P., Inza, I., Sierra, B.: Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence, *accepted for publication* (2004)
19. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning, Vol. 46 (1-3). (2002) 389–422
20. Jaeger, J., Sengupta, R., Ruzzo, W.L.: Improved gene selection for classification of microarrays. Proc. of Pacific Symposium on Biocomputing, (2003) 53–64
21. Qi, H.: Feature selection and kNN fusion in molecular classification of multiple tumor types. Proc. of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, (2002)
22. Hanczar, B., Courtine, M., Benis, A., Hennegar, C., Clément, K., Zucker, J.D.: Improving classification of microarray data using prototype-based feature selection. ACM SIGKDD Explorations Newsletter, Vol. 5 (2). (2003) 23–30
23. Zheng, G., Olusegun, E., Narasimhan, G.: Neural network classifiers and gene selection methods for microarray data on human lung adenocarcinoma. Prof. of Critical Assessment of Microarray Data Analysis, (2003) 63–67
24. Hochreiter, S., Obermayer, K.: Feature selection and classification on matrix data: from large margins to small covering numbers. Advances in Neural Information Processing Systems, Vol. 15. (2003) 913–920
25. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature selection for SVMs. Advances in Neural Information Processing Systems, Vol. 13. (2001) 668–674
26. Pal, S., Shiu, S.: Foundations of Soft Case-Based Reasoning. John Wiley, New York (2004)
27. Pal, S., Mitra, P.: Case Generation Using Rough Sets with Fuzzy Representation. IEEE Transactions on Knowledge and Data Engineering, Vol. 16 (3). (2004) 292–300
28. Riesbeck, C.K., Schank, R.C.: Inside Case-Based Reasoning, Lawrence Erlbaum Associates, Hillsdale, NJ, US (1999)
29. Fdez-Riverola, F., Corchado, J.M.: FSfRT, Forecasting System for Red Tides. An Hybrid Autonomous AI Model. Applied Artificial Intelligence, Vol. 17 (10). (2003) 955–982
30. Pal, S.K., Dilon, T.S., Yeung, D.S.: Soft Computing in Case Based Reasoning, Springer Verlag, London (2000)
31. Sankar, K.P., Simon, C.K.S: Foundations of Soft Case-Based Reasoning, Wiley-Interscience, Hoboken, New Jersey (2003)
32. Fdez-Riverola, F., Corchado, J.M.: Employing TSK Fuzzy models to automate the revision stage of a CBR system. Current Topics in Artificial Intelligence, LNAI 3040. (2004) 302–311
33. Gutierrez, N.C., López-Pérez R., Hernández, J.M., Isidro, I., González, B., García, J.L., Ferminán, E., Lumbreras, E., San Miguel, J.F.: Gene expression profile reveals deregulation of new genes with relevant functions in the different subclasses of acute myeloid leukemia. Blood, Vol. 102 (11). (2003)
34. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. of the National Academy of Sciences of the United States of America, Vol. 99(10). (2002) 6561–6572
35. Aaronson, J.S., Juergen, H., Overton, G.C.: Knowledge Discovery in GENBANK. Proc. of the First International Conference on Intelligent Systems for Molecular Biology, (1993) 3–11