GENE-CBR: A CASE-BASED REASONIG TOOL FOR CANCER DIAGNOSIS USING MICROARRAY DATASETS*

FERNANDO DÍAZ

Escuela Universitaria de Informática, University of Segovia, Spain

FLORENTINO FDEZ-RIVEROLA

Escuela Superior de Ingeniería Informática, University of Vigo, Spain

JUAN M. CORCHADO
Facultad de Ciencias, University of Salamanca, Spain

Gene expression profiles are composed of thousands of genes at the same time, representing the 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 design methods capable of overcoming current limitations of state-of-the-art algorithms is crucial to the development of successful applications. This paper presents GENE-CBR, a hybrid model that can perform cancer classification based on microarray data. The system employs a case-based reasoning model that incorporates a set of fuzzy prototypes, a growing cell structure network and a set of rules to provide an accurate diagnosis. The hybrid model has been implemented and tested with microarray data belonging to bone marrow cases from 43 adult patients with cancer plus a group of six cases corresponding to healthy persons.

Key words: microarray datasets, cancer diagnosis, gene-CBR, hybrid model, fuzzy patterns, gene selection.

1. INTRODUCTION AND MOTIVATION

In recent years DNA microarray technology has become a fundamental tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease (Schena *et al*. 1995). 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 hybridized to ESTs (Lipshutz *et al*. 1999). Since the number of examined genes in an experiment runs to the thousands, different data mining techniques have been intensively used to analyze and discover knowledge from gene expression data (Piatetsky-Shapiro and Tamayo 2003). 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.

* This work was supported in part by the MCYT project TIC2003-07369-C02-02.

Address correspondence to Florentino Fdez-Riverola, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense. E-mail: riverola@uvigo.es

For several years we have been working in the identification of techniques to automate the reasoning cycle of several complex case-based reasoning (CBR) systems applied to complex domains (Fdez-Riverola *et al*. 2005; Corchado *et al*. 2003; Corchado *et al.* 2004). Based on our previous research work, we are interested in the development of a robust CBR 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. The hybrid system proposed in this paper presents a new synthesis that brings several artificial intelligence subfields together (fuzzy sets, artificial neural networks and *if-then* rule-sets). The retrieval, reuse, revision and learning stages of the CBR system use these techniques to facilitate the CBR adaptation to the domain of biological discovery with microarray datasets.

GENE-CBR is a model that can perform cancer classification based on microarray data. In order to store the information belonging to each sample, the system uses a fuzzy codification to represent the gene expression levels of each sample. This operation permits the generalization over the whole case base in order to tackle intra-experimental and inter-experimental variations in the data. Based on the fuzzy discretization of real gene expression data into a small number of fuzzy membership functions, the system is capable of constructing a set of prototypes that are able to represent the main characteristics of previously ascertained classes.

GENE-CBR employs a case-based reasoning model that incorporates a set of fuzzy prototypes for the retrieval of relevant genes, a growing cell structure (GCS) network and a proportional weighted voting algorithm for the clustering of similar patients and the assignation of an initial class, and a set of See5 rules used to formalize the knowledge extraction to justify the results. In the experiments reported in this paper, we work with a database of bone marrow cases from 43 adult patients with Acute Myeloid Leukemia (AML) plus a group of six samples belonging to healthy persons for test purposes. 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.

The paper is organized as follows: Section 2 presents previous related research work in applying CBR to genetic/bioinformatics domains. Section 3 discusses our proposal in detail, showing how GENE-CBR stores the available data and explaining how to implement each phase of the hybrid system. Section 4 introduces the experimental test bed carried out, discussing and analyzing the results obtained from different available data. Finally, Section 5 concludes summarizing the results obtained and the advantages of our proposed model.

2. RELATED WORK

As mentioned earlier, the development of the DNA microarray technology has the power to generate enormous amounts of data, but it also requires sophisticated analysis techniques in order to be able to extract biologically meaningful knowledge from the raw data.

The fields of data mining (DM) and machine learning (ML) have been dealing with these kinds of tasks for a few decades. However, the data from microarrays has some features that make it hard to analyze. In recent years, ML and DM techniques have been used adequately enough to address the incipient research areas that have arisen in DNA microarray analysis (Piatetsky-Shapiro and Tamayo 2003).

In this context, case-based reasoning is particularly applicable to this problem domain because it (*i*) supports a 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 (Jurisica and Glasgow 2004). CBR systems can be used to propose new solutions or evaluate solutions to avoid potential problems. In the work of Aaronson *et al*. (1993) it is suggested that analogical reasoning is particularly applicable to the biological domain, partly because biological systems are often homologous (rooted in evolution). Moreover, 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.

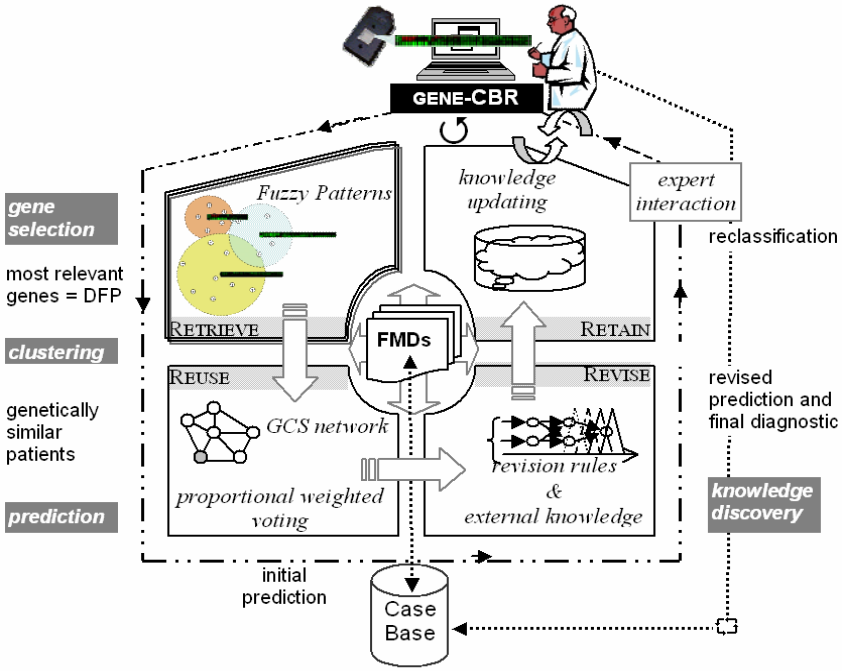
In this sense, the work of Arshadi and Jurisica (2005) proposes a mixture of experts for case-based reasoning (MOE4CBR) developing a method that combines an ensemble of CBR classifiers with spectral clustering and logistic regression. This approach not only achieves higher prediction accuracy, but also leads to the selection of a subset of features that have meaningful relationships with their class labels. Previously, Costello and Wilson (2003) showed their initial work in applying a CBR approach to the problem of gene-finding in mammalian DNA. The results obtained from their experiments indicate that it is certainly feasible to do DNA-to-DNA comparisons in order to isolate relevant coding regions. A previous successful work in the same area using CBR was carried out by Shavlik (1991).

Lieber and Bresson (2000) showed how their CASIMIR/CBR system was able to suggest solutions for breast cancer treatment by adapting the rules of a previous rule-based system (CASIMIR/RBR). Jurisica and Glasgow (2004) demonstrate how case-based reasoning can be applied to assist in analyzing genomic sequences and determining the structure of proteins. They also provide an overview of several other applications in molecular biology that have benefited from case-based reasoning.

3. THE MODEL

Case-based reasoning is a computational reasoning paradigm that involves the storage and retrieval of past experiences to solve new problems (Watson 1997). An advantage of CBR systems as a problem-solving paradigm is that it is applicable to a wide range of problems, and is particularly relevant in scientific domains, where there is a wealth of data but often a lack of theories or general principles. In this case, we have integrated data and knowledge in the GENE-CBR system to facilitate the classification of patients with AML and the identification of the group to which a given patient may belong. The system solution is justified and new knowledge is incorporated into the system with each patient.

In order to initially construct the model case base starting from the available patients data (lower part of Figure 1), GENE-CBR stores the gene expression levels of each sample in its case base. The system always deals with a fuzzy codification of the values stored. During the retrieval stage, the original case vectors are transformed into a fuzzy microarray descriptors (FMD). Each FMD is a comprehensible descriptor of the sample in terms of a linguistic label for each gene expression level (central part of Figure 1). This transformation is carried out by mean of a fuzzy discretization



process as detailed in the following subsections.

FIGURE 1. GENE-CBR architecture.

Based on the FMD representation created from the case base, a set of fuzzy patterns (FP) is constructed that represents the main characteristics of the *a priori* known classes (left top square in Figure 1). Each class in the system is then represented by a FP that holds the fuzzy codification of gene expression levels for those genes that were flagged as relevant for this class. Several FPs are generated from the data in a supervised way, each one representing a group of FMDs for each specific pathology.

The retrieval stage in our GENE-CBR system uses the FPs in order to select the most representative genes given a new patient. This phase can be thought of as a gene selection step, in which the aim is to retrieve the list of genes that might be most informative given a new sample to classify. Since it is highly unlikely that all the 22,283 genes have significant information related to cancer classification and the dimensionality would be too great if all the genes were used, it is necessary to explore an efficient way to obtain the most suitable group of genes. In order to make this selection, our GENE-CBR system selects those fuzzy patterns from its case base which

are the nearest to any new case obtained. Then, for each one of the selected FPs, the GENE-CBR system computes its associated DFP (a pattern which only includes the genes that are necessary in order to discriminate the novel instance from other different classes). Finally, the selected genes for the new case are obtained by joining together the genes belonging to the DFPs considered.

The adaptation of previous cases in order to solve a new FMD is accomplished in the reuse stage (left bottom square in Figure 1). A growing cell structures network is trained with the whole case base, only taking the existing cases represented by the genes selected in the previous stage as input. Then, the new FMD is presented to the GCS network and the patients most similar from a genetic point of view are retrieved. Based on this grouping, a proportional weighted voting mechanism is applied that ponders the level of similarity with the new FMD. An initial class is assigned by the GENE-CBR system from among the existing pathologies.

In the revision stage (right bottom square in Figure 1) the expert is provided with useful data about the decision made by the system. This information contains the selected DFP genes, the grouping made by the GCS network, the weighting assigned to each class and a set of See5 (Quinlan 2000) classification rules generated from the most similar patients. The expert contrasts the initial prediction given by the system with other external information such as patient karyotype or clinical history in order to ascertain a revised prediction and a final diagnostic.

Every time a new FMD is solved, the internal structure of the GENE-CBR system is updated (right top square in Figure 1). The new FMD is associated to its corresponding class and added to the case base. The affected FP is updated and the system marks the most similar patients selected for future classifications. In this stage the GENE-CBR system changes to edit-mode and the expert is permitted to update patients classification taking into account the new knowledge generated.

3.1. Case Representation

Input space reduction is often the key phase in the building of an accurate classifier (Cakmakov and Bennani 2003). Based on the fuzzy discretization method presented below, the proposed system is able to represent any microarray by means of its fuzzy microarray descriptor (FMD). This descriptor is a comprehensible description for each gene in terms of one from the following linguistic labels: LOW, MEDIUM and HIGH. Obviously, this process implies a loss of information, but this representation is representative from a medical point of view because practitioners are interested in which genes are over-expressed or under-expressed, rather than the exact level of expression in each gene. Therefore, the representation proposed fits the requirements of the system users. Moreover, from a set of FMDs the system is also able to construct a prototype, known as a fuzzy pattern (FP), which characterizes and summarizes the most relevant values of expression levels within this set of samples. As shown in Subsection 3.2, below, this ability within the system will be exploited in order to select a subset of relevant genes from these fuzzy patterns.

The rest of this subsection describes the discretization schema and the construction of a fuzzy pattern from samples, as a basic mechanism of the proposed system in order to represent the information available.

3.1.1. Fuzzy Microarray Descriptor (FMD)

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 those proposed by Pal *et al*. (2000). These authors suggest a useful polynomial function that approximates a Gaussian membership function, where its centre and amplitude depend on the mean and 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. A graphical representation of the considered membership functions is shown in Figure 2.

The values of the centers and amplitudes of each membership function can be computed from the original data set $\boldsymbol{D}$. In terms of the case base of the proposed system (see Figure 1) an element of the data set $\boldsymbol{D}$ represents a microarray. For each probe (associated with each EST tested by the microarray), the values of the parameters of the centre and amplitude for each linguistic label can be pre-computed. This information must be part of the case base, concretely it must be stored in the FMD layer. Therefore, the FDM layer acts as the interface between each module of the system (retrieve, reuse, revision and retention modules) and the warehouse of the raw data given by the microarrays available.

Once the three membership functions for each feature F_j have been defined, 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 labels defined - LOW, MEDIUM and HIGH. The discriminatory criterion for any label is simply defined by:

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

As shown in Figure 2, 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 2). From this fact, it must be interpreted that 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 two linguistic labels simultaneously, since at the significance level given by Θ, any assignment of the measure to a linguistic label is significant (see, the shadow region of the intersection of label MEDIUM and HIGH in Figure 2).

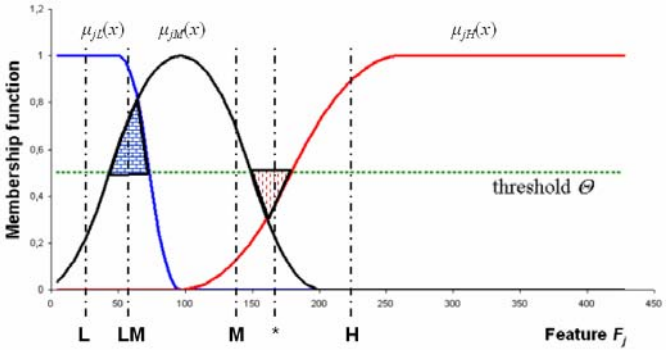


FIGURE 2. Membership functions for the linguistic labels: LOW, MEDIUM and HIGH.

After this, any observation $\boldsymbol{x}$ belonging to $\boldsymbol{D}$, and originally characterized by n numeric features $\boldsymbol{F} = \{F_1, ..., F_n\}$, namely $\boldsymbol{x} \in \mathbf{R}^n$, will be transformed into a new vector with $3n$ binary features ($\boldsymbol{x'} \in \{0, 1\}^{3n}$). Therefore, the real value of feature F_j for the observation $\boldsymbol{x}$, denoted by x_j, is replaced by the three binary values given by Expression (1) for each linguistic label, that is to say, by the tuple $\langle F'_{jL}(x_j), F'_{jM}(x_j), F'_{jH}(x_j)\rangle$. In order to reduce the dimension of the transformed data, the possible status with respect to the three linguistic labels are unified into only one label. Given that the three binary values for each feature are replaced by a single label, $F''_j(\boldsymbol{x}) \in \{\text{L, LM, M, MH, H, *}\}$. If only one of the three binary values is active, the respective label is assigned: L (LOW), M (MEDIUM), and H (HIGH). As mentioned above, 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 unifying criteria is given completely by Expression (2).

$$F''_j(\boldsymbol{x}) = \begin{cases} L & \text{if } F'_{jL}(x_j) = 1 \wedge F'_{jM}(x_j) = 0 \wedge F'_{jH}(x_j) = 0 \\ LM & \text{if } F'_{jL}(x_j) = 1 \wedge F'_{jM}(x_j) = 1 \wedge F'_{jH}(x_j) = 0 \\ M & \text{if } F'_{jL}(x_j) = 0 \wedge F'_{jM}(x_j) = 1 \wedge F'_{jH}(x_j) = 0 \\ MH & \text{if } F'_{jL}(x_j) = 0 \wedge F'_{jM}(x_j) = 1 \wedge F'_{jH}(x_j) = 1 \\ H & \text{if } F'_{jL}(x_j) = 0 \wedge F'_{jM}(x_j) = 0 \wedge F'_{jH}(x_j) = 1 \\ * & \text{if } F'_{jL}(x_j) = 0 \wedge F'_{jM}(x_j) = 0 \wedge F'_{jH}(x_j) = 0 \end{cases}. \quad (2)$$

The resulting labels for each gene of a microarray $\boldsymbol{x}$ define the FMD for this microarray and the FMD layer is responsible for computing the fuzzy descriptor. As mentioned above, its operation is governed by the available data (the computed FMD depends on the available microarrays via the pre-computed centers and amplitudes for each feature) and provides a higher-level representation (a knowledge representation) of the raw data. The knowledge representation given by the FMD layer can also be tuned by the user via the parameter Θ, which is introduced in Expression (1). As the value of this parameter goes to 1, the computed values of the FMD will be more spe-

cific (and, simultaneously, less general). Therefore, the knowledge representation given by the FMDs is graded by the parameter Θ.

The input parameter *context* represents the set of parameters (the centre and amplitude) for the membership functions associated with each gene array probe. These values are computed from the whole set of microarrays available in the case base. The parameter Θ is the threshold value which controls the activation of a linguistic label.

3.1.2. Fuzzy Pattern

A fuzzy pattern (FP) is a higher concept constructed from a set of FMDs. A fuzzy pattern can be viewed as a prototype of the set of FMDs from which it is constructed. Therefore, the fuzzy pattern can capture relevant and common information about the gene expression levels of these FMDs. Obviously, this fact can be of interest, if the set of initial observations are labeled with the same kind of cancer. A fuzzy pattern can be constructed as follows. Given a subset of observations $\boldsymbol{D}_i = \left\{ \boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, ..., \boldsymbol{x}_{i_m} \right\} \subseteq \boldsymbol{D}$, which have associated the same class label C_i, for any observation $\boldsymbol{x}_{i_l}$ $(i_1 \leq i_l \leq i_m)$, their FMD can be computed as indicated in the previous subsection. Therefore, for each observation, $\boldsymbol{x}_{i_l} \in \boldsymbol{D}_i$, its associated FMD will have for each feature $F_j^{"}(\boldsymbol{x}_{i_l})$ a label from the set {L, LM, M, MH, H, *} according to the Expression (2). The fuzzy pattern (corresponding to the class C_i) is constructed from the FMDs of each one of the initial observations, selecting those labels of features which are different to the label "*" and have a relative appearance frequency in set $\boldsymbol{D}_i$ equal to or greater than a predefined ratio Π (0 < Π ≤ 1, for example, Π = 2/3).

Formally, for each feature F_j, the appearance frequency of any label $E \in \boldsymbol{E}$ = {L, LM, M, MH, H, *} in the set $\boldsymbol{D}_i$, $\pi_{ij}(E)$, can be computed according to the expression given by

$$\pi_{ij}(E) = \frac{\sum_{i_1 \leq i_l \leq i_m} \delta_j(\boldsymbol{x}_{i_l}, E)}{i_m}, \text{ where } \delta_j(\boldsymbol{x}_{i_l}, E) = \begin{cases} 1 & \text{if } F"_j(\boldsymbol{x}_{i_l}) = E \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

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:

$$\boldsymbol{P}_i = \left\{ \bigwedge_{F"_j \in \boldsymbol{F}"} \left\langle F"_j, E^j, \pi^j \right\rangle : E^j = \arg\max_{E \in \boldsymbol{E}} \left\{ \pi_{ij}(E) \right\} \wedge E^j \neq * \wedge \pi^j = \pi_{ij}(E^j) \geq \Pi \right\}. \quad (4)$$

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 the 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 transformation into

FMD of an initial observation. Moreover, 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.

The interest of a pattern is greater if the microarrays included in the list are members of a target concept (for example, a specific kind of cancer). The input parameter *context* represents the set of parameters (the centre and amplitude) for the membership functions associated with each probe of the gene array. This context is computed from the whole set of available microarrays (not only those belonging to a target concept). Finally, the input parameters Θ and Π control the FMD transformation of a microarray and the labels which are included in the FP, respectively. The computed fuzzy pattern is returned in the output parameter *FP*.

3.2. Retrieval and Gene Selection

The goal of the retrieve stage is, given a problem case (new microarray), to select a reduced number of genes which are relevant for classifying the new microarray into one of the classes of cancer represented in the case base. In order to perform this operation, the system computes the fuzzy distance of the new instance to the existing fuzzy prototypes and then selects the nearest ones to the new problem case. From each one of these prototypes, the system computes a unique discriminant fuzzy pattern (DFP). This DFP is computed as the common genes of the initial FPs which, moreover, are discriminant among two different classes of cancer. The union of the genes belonging to each DFP will be the genes that the system returns as the relevant genes in order to classify the new microarray and, therefore, it will be the output of the retrieve stage. A major issue of the proposed model is its ability to determine the similarity among patterns (or FMDs).

In order to explain how we calculate this relation we need to previously define the similarity between linguistic labels (represented by 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 (see Fdez-Riverola *et al*. (2005) for details). In this way, 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 3.

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 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. Moreover, the values of similarity among membership functions for the same feature can be pre-computed at the same time that the parameters of these functions (centers and amplitudes) are determined, and therefore, can be stored by the FMD layer to reduce the computational requirements.

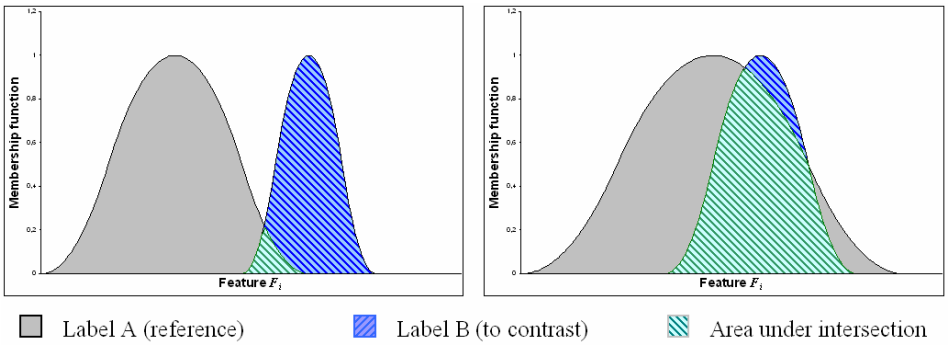


FIGURE 3. Relationship between the area below the membership function and the similitude of linguistic labels.

Now, we are interested in measure the distance of a new observation $\boldsymbol{x}$ to an specific fuzzy pattern $\boldsymbol{P}$ of the form:

$$\boldsymbol{P} = \bigwedge_{F''_j \in \boldsymbol{F''}} \left\langle F''_j, E^j, \pi^j \right\rangle = \left\langle F''_{j_1}, E^{j_1}, \pi^{j_1} \right\rangle \wedge \ldots \wedge \left\langle F''_{j_n}, E^{j_n}, \pi^{j_n} \right\rangle . \tag{6}$$

where j_n is the number of variables which the pattern has. After computing the FMD of the new observation $\boldsymbol{x}$, the distance between $\boldsymbol{x}$ and the pattern $\boldsymbol{P}$, denoted by $d(\boldsymbol{P}, \boldsymbol{x})$, is defined as:

$$d(\boldsymbol{P}, \boldsymbol{x}) = \frac{j_n}{\sum_{j_1 \le j_k \le j_n} sim(E^{j_k}, F''_{j_k}(\boldsymbol{x})) \cdot \pi^{j_k}} - 1 . \tag{7}$$

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).

Once the distance between a FMD $\boldsymbol{x}$ to a pattern $\boldsymbol{P}$ can be measured, at the retrieve stage the distance of the new case $\boldsymbol{x}$ to all available patterns represented in the case base of the GENE-CBR system is computed in order to select the nearest patterns to $\boldsymbol{x}$. Since each pattern is representative of a collection of microarrays belonging to the same class, we can assume that the genes included in a pattern, are significant to the classification of any novel instance within the class associated with that pattern. Moreover, if several patterns are selected as possible matches for the new case, we are interested in those genes that allow us to discriminate the new case from one class with regard to the others. Here we introduce the notion of discriminant fuzzy pattern (DFP) with regard to a collection of fuzzy patterns. Briefly, given a collection of fuzzy patterns, from each FP it can be computed its DFP as a projection of the original FP in such a way that only includes within the DFP those genes which are present

in any other pattern of the collection but with a different label. Therefore, the DFP version of a FP only includes those genes that can serve to differentiate it from the rest of the patterns.

3.3. Reuse and Initial Prediction

Each time a new problem case (microarray) is presented to the system, a subset of relevant genes is created as explained in Subsection 3.2. This subset includes the genes that are defined as most relevant for the incoming patient. The next step that GENE-CBR accomplishes is the grouping of those cases that are most similar but only taking into account the genetic information provided by the selected genes. In order to do this, our system incorporates a growing cell structure network.

To illustrate the working model of the GCS network inside the GENE-CBR system, a two-dimensional space is used, where the cells (neurons) are connected and organized into triangles (Fritzke 1993). Each cell in the network is associated with a weight vector, *w*, of the same dimension as the number of relevant genes selected in the previous step (size of the DFP array). At the beginning of the learning process, the weight vector of each cell is initialized with random values (Fritzke 1993). The basic learning process in a GCS network consists of topology modification and weight vector adaptations carried out.

It is important to highlight that the final goal of the GCS network is to cluster all patients that are genetically similar given a selected group of genes (DFP) and without taking into account their previous assigned classes. Our proposed method aims to find new relations between the patients even now unknown. Therefore, it is possible and not contradictory to group together patients suffering different (but genetically related) diseases. The automatic discovery of unknown relevant knowledge from cases is particularly useful in CBR schemes based on the *k*-NN algorithm (Dasarathy 1991), where *k* is typically constant. The proposed GCS scheme does not rely on a predefined, fixed *k*, rather that the set of retrieved patients depends on the groupings of cognate FMDs in the case base. Thus, this stage could be thought of as a *dynamic k-nearest neighbour* scheme. Since such networks contain explicit distance information, they can be used effectively in CBR to represent an indexing structure which indexes sets of cases in the case base and a similarity measurement between case sets (Azuaje *et al*. 2000).

Every time a new FMD is presented to GENE-CBR, the retrieval step of our hybrid system filters the most relevant genes in order to construct the DFP. This group of genes is then employed to train the GCS network with the whole case base. From the trained CGS network, a new sorted vector of pairs holding the similarity of each selected patient with the new case, *S*, is generated.

In order to apply the proportional weighted voting schema, we need to ponder the vote of each patient contained in vector *S*. To do this, a weight α_j for each retrieved patient, k_j, is calculated based on the position (*pos*) that it occupies in the vector *S* and the level of similarity with the target case, Sim_j. For this task, Expression (8) is used.

$$\alpha_j = Sim_j \frac{2^{|S|-1}}{(2^{|S|}-1)2^{pos-1}}. \tag{8}$$

Therefore, the initial prediction made by GENE-CBR when a target case is presented to the system depends on both the number of selected patients (those genetically similar taking into account the genes belonging to the DPF pattern) and the level of similarity with the target case. The solution proposed by the system is the class corresponding to the disease with the highest score.

As we can surmise, it is easy to introduce a rejection mechanism in the voting model. We simply use a threshold *T* to indicate whether the score received by the best matching class is sufficiently strong (passing quota). In the event that the score received by the matching class is less than *T*, then the target case remains unclassified.

3.4. Revision and Knowledge Extraction

Every time GENE-CBR system evolves to the previous stage, one of the two following situations will be reached: (*i*) the system is able to classify the target case (so the assigned score exceeded the threshold *T*) or (*ii*) the system can propose a classification but it is not as accurate as the one required (i.e. a new subtype of cancer could be detected). Any of these situations is reported to the expert that manually reexamines the proposed solution in the light of the justification given by the system. At this stage, the system provides the expert with useful data about the decision made by GENE-CBR. This information contains the selected DFP gene values for the most similar patients, the grouping made by the GCS network, the weight assigned to each class and a set of See5 classification rules (Quinlan 2000) generated starting from the most similar patients. The set of See5 rules are used to formalize knowledge extraction and severs as result justification. Starting from the set of previously generated rules, a causal network representation can be extracted in order to construct a connected graph for further analysis.

Each See5 rule consists of a rule number with the purpose of identification. Following the rule number, some statistical information is presented in the form of (*n*, lift *x*) or (*n*/*m*, lift *x*) summarizing the performance of the rule. Similarly to a leaf, *n* is the number of training cases covered by the rule and *m*, if it appears, shows how many of them do not belong to the class predicted by the rule. The rule's accuracy is estimated by the Laplace ratio $(n-m+1)/(n+2)$. The lift *x* is the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set. The rule is composed of one or more conditions that must all be satisfied if the rule is to be applicable and a class predicted by the rule. Finally, the rule shows a value between 0 and 1 indicating the confidence with which this prediction is made.

Using this information, the expert contrasts the initial prediction given by the system with other external information such as patient karyotype or clinical history in order to ascertain a revised prediction and a final diagnostic. This external information is automatically shown to the expert by GENE-CBR because it is previously linked to each existing case by a Knowledge Acquisition Module (KAM).

3.5. Knowledge Acquisition

One of the main goals of microarray data analysis is discovery of biological knowledge. Every time GENE-CBR system executes the previous stage, the expert is provided with new knowledge and a final prediction is generated. The continuous

learning in the system is achieved in two ways with the help of an expert. Firstly, a new case containing the microarray data and the assigned class is automatically stored in the case base for future use. Secondly, the expert can reconsider previous classifications in the light of new evidence.

In order to help the expert acquire new information about the problem domain, GENE-CBR implements several tools for inspecting the knowledge base of the system. In this sense, the most interesting issues from a biological perspective are: (*i*) the subset of genes finally selected for the definition of the DFP pattern, that could help to reduce the dimensionality of gene expression data and the search for relevant (explicative) genes (*ii*) the clustering generated by the GCS network, that could be a good indicator of the occurrence of new classes or new subtypes of an existing one, (*iii*) the similarity (coherence) among the retrieved patients of each cluster, that aids in the definition of a sound set of characteristics that better represents each class, and (*iv*) the set of See5 rules, that can be viewed as general knowledge summarizing the relevance of the acquired knowledge.

An especially interesting goal in our research team is the identification of supergenes, that is, genes that could have extra potential in classifying different types of cancer. Those genes satisfy an important characteristic: they serve as predictors for their class and differ from other classes.

4. RESULTS

Bone marrow (BM) samples from 43 adult patients with newly de novo diagnosed AML were analyzed. All samples contained more than 80% blast cells. The median age was 36 years (range 14-70 years). Patients were classified according to the WHO classification into 4 subgroups: a) 10 APL with t(15;17) confirmed by FISH studies with LSI PML/RARA probe (Vysis, Stuttgart, Germany), b) 4 AML with inv(16) confirmed by FISH analysis with LSI CBFB probe (Vysis); c) 7 acute monocytic leukemias and d) 22 non-monocytic AML without recurrent cytogenetic translocations. Six BM samples from healthy donors were also included as a control group (see Table 1 for the distribution of samples). Each case (microarray experiment) stores 22,283 ESTs corresponding to the expression level of thousands of genes. The data consisted of 1,091,867 scanned intensities.

TABLE 1. Distribution of samples taking into account the type of cancer.

	Healthy	APL	AML-inv()	AML-mono	AML-other
Number of patients	6	10	4	7	22

The goal of this study is to evaluate the predictive accuracy of the model proposed, namely, the ability of the model for classifying new samples in the pre-defined groups. Due to the low numbers of samples, a cross-validation strategy has been chosen to estimate the accuracy of the model (specifically a 4 fold cross-validation, considering the number of samples in group AML-Inv).

Figure 4 shows the cross-validation error for different values of the parameters Θ and Π. It is used 4-fold cross-validation, dividing the set of samples at random into four approximately equal-sized parts. The four parts were roughly balanced, ensuring

that the classes were distributed proportionally among each of the four parts. 4-fold cross-validation works as follows: we fit the model on 75% of the samples and then predict the class labels of the remaining 25% (the test samples). This procedure is repeated four times, with each part playing the role of test sample and the errors on all four parts averaged out to compute the overall error.

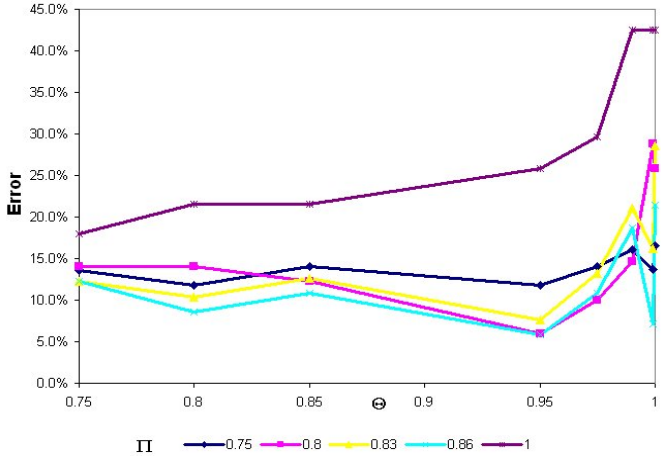


FIGURE 4. Variation of the 4-fold cross-validation error depending on the model parameters.

Figure 4 shows the results from the rank 0.75 to 1 (x axis) for the parameter Θ (the threshold which establishes if a linguistic label is assigned to the expression level of a gene within a sample) and different values (each one of the lines in graph) of the parameter Π (the frequency threshold which determines if the expression level of a gene is included in the fuzzy pattern of the class considered, depending on its frequency of appearance in the available samples of such a class). The minimum error, 7.63%, is reached at values 0.95 and 0.86, for the parameter Θ and parameter Π, respectively. Parameter Θ allows us to modulate the representation of a microarray in the form of a fuzzy microarray descriptor (FMD) as shown in Subsection 3.1.1, whereas parameter Π determines the ability of the system to construct fuzzy patterns (see Subsection 3.1.2).

The results of the GENE-CBR system have been compared with the ones obtained by the PAM (*Prediction Analysis of Microarrays*) technique (Tibshirani 2002). This technique is considered as a reference technique in the field of the microarray analysis and is based on an enhancement of the simple nearest prototype (centroid) classifier. The method also identifies subsets of genes that best characterize each class. The first analysis leaves out the healthy patients and we initially compare the genes selected by the GENE-CBR with regard to the ones selected by PAM software. The GENE-CBR is able to select 19 different genes which belong to all the discriminant fuzzy patterns computed by the GENE-CBR system. PAM software selects a subset of 382 significant genes (the complete list is not shown due to its large size). The shrinkage parameter Δ of the PAM algorithm was 2.5, since beyond this value the 4-fold cross-validation error increased considerably. Comparing the two lists of genes it can be observed that all the genes selected by the GENE-CBR system are also selected by

the PAM software. That is, the genes selected by our system are a subset of the genes selected by PAM and, therefore they are significant genes (at least to the same degree that PAM determines them, and possibly with a higher degree of significance due to the reduced number of genes).

With respect to the accuracy of the classification, Table 2 summarizes the results of both systems in order to classify a set of novel instances. The followed strategy was a 4-fold cross-validation in order to estimate the accuracy of each system. The results also show the standard error (in a smaller font than the average error) of the mean accuracy as a metric of the variability of the estimated accuracy. After performing a t-test it can be stated that differences exist between the estimated accuracy of the two systems. This difference is statistically significant, but not very significant (the p-value is 0.041).

TABLE 2. Estimated accuracy of the GENE-CBR system and the PAM software.

	Training	4-fold cross-validation (only patients with cancer)	4-fold cross-validation (all the samples)
GENE-CBR	0%	5.98 % ± 2.03 %	7.76 ± 2.17
PAM	0%	12.42 % ± 1.41 %	17.59 ± 1.50

The GENE-CBR system improves its response with regard to the PAM technique when the system also needs to classify healthy patients. In this case, the number of genes selected by the GENE-CBR system increases to 59 genes, whereas the genes selected by PAM total about 500 genes. Moreover, the difference between the mean accuracy is statistically very significant since the p-value is less than 0.01.

5. CONCLUSIONS

This paper presents the successful implementation of a CBR system for cancer diagnosis using microarray datasets. The CBR systems proposed support rich and evolvable representations of experiences, solutions and feedback. The biologist often uses a form of reasoning similar to CBR, whereby experiments are designated and performed based on the similarity between features of a new situation and those of known experiences. GENE-CBR allows the use of combined techniques that can be used for gene selection, clustering, knowledge extraction and prediction.

The results obtained from the empirical studies carried out are very promising and they back up the idea that CBR systems can offer a number of advantages in the microarray domain. Concretely, GENE-CBR allows us to obtain a more general knowledge about the problem domain and to gain a deeper insight into the importance of each gene related to each pathology. Moreover, we show how our GENE-CBR system outperforms the results obtained by specific classification techniques such as PAM, when applied to our AML experimental data. Moreover, statistical tests carried out show that this difference is significant.

REFERENCES

AARONSON, J. S., JUERGEN, H., and OVERTON, G. C. 1993. Knowledge Discovery in GENBANK. *In* Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, pp. 3–11.

ARSHADI, N., and JURISICA, I. 2005. Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains. IEEE Transactions on Knowledge and Dat Engineering, 17(8):1127-1137.

AZUAJE, F., DUBITZKY, W., BLACK, N., and ADAMSON, K. 2000. Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach. IEEE Transactions on Systems, Man and Cybernetics, B, 30(3):448–460.

CAKMAKOV, D., and BENNANI, Y. 2003. Feature selection for pattern recognition, Informa Press.

CORCHADO, J. M., AIKEN, J., CORCHADO, E., LEFEVRE, N., and SMYTH, T. 2004. Quantifying the Ocean's CO2 Budget with a CoHeL-IBR System. *In* Proceedings of the 7th European Conference on Case-based Reasoning, pp. 533–546.

CORCHADO, J. M., CORCHADO, E. S., AIKEN, J., FYFE, C., FDEZ-RIVEROLA, F., and GLEZ-BEDIA, M. 2003. Maximum Likelihood Hebbian Learning Based Retrieval Method for CBR Systems. *In* Proceedings. of the 5th International Conference on Case-Based Reasoning, pp. 107–121.

COSTELLO, E., and WILSON, D. C. 2003. A Case-Based Approach to Gene Finding. *In* Proceedings of the Fifth International Conference on Case-Based Reasoning Workshop on CBR in the Health Sciences, pp. 19-28.

DASARATHY, V. 1991. Nearest Neighbor (NN) Norms NN pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA.

FDEZ-RIVEROLA, F., DÍAZ, F., BORRAJO, M. L., YÁÑEZ, J. C., and CORCHADO, J. M. 2005. Improving gene selection in microarray data analysis using Fuzzy Patterns inside a CBR system. *In* Proceedings of the 6th International Conference on Case-Based Reasoning, pp. 191–205.

FRITZKE, B. 1993. Growing Cell Structures - A Self-organizing Network for Unsupervised and Supervised Learning. Technical Report, International Computer Science Institute, Berkeley.

JURISICA, I., and GLAWGOW, J. 2004. Applications of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics, 25(1):85–95.

LIEBER, J., and BRESSON B. 2000. Case-Based Reasoning for Breast Cancer Treatment Decision Helping. *In* Proceedings of the 5th European Workshop on Case-Based Reasoning, pp. 173–185.

LIPSHUTZ, R. J., FODOR, S. P. A., GINGERAS, T. R., and LOCKHART, D. H. 1999. High density synthetic oligonucleotide arrays. Nature Genetics, 21:20–24.

PAL, S. K., DILON, T. S., and YEUNG, D. S. 2000. Soft Computing in Case Based Reasoning, Springer Verlag, London.

PIATETSKY-SHAPIRO, G., and TAMAYO, P. 2003. Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter, 5(2):1–5.

QUINLAN, J. R. 2000. Data mining tools See5 and C5.0. http://www.rulequest.com/see5-info.html/.

SCHENA, M., SHALON D., DAVIS, R., and BROWN, P. O. 1995. Quantitative monitoring of gene expression patterns with a cDNA microarray. Science, 270:467–470.

SHAVLIK, J. 1991. Finding Genes by Case-Based Reasoning in the Presence of Noisy Case Boundaries. *In* Proceedings of the DARPA Workshop on Case-Based Reasoning, pp. 327–338.

TIBSHIRANI, R. HASTIE, T., NARASIMHAN, B., and CHU, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *In* Proceedings of the National Academy of Sciences, 99(10):6567–6572.

WATSON, I. 1997. Applying Case-based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann, San Mateo, CA.