Chapter 13
Computational Intelligence Techniques for Classification in Microarray Analysis

Juan F. De Paz[1], Javier Bajo[2], Sara Rodríguez[1], and Juan M. Corchado[1]

[1] Departamento de Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, España
[2] Universidad Pontificia de Salamanca
Compañía 5, 37002, Salamanca, España
{fcofds,srg,corchado}@usal.es, jbajope@upsa.es

Abstract. During the last few years there has been a growing need for using computational intelligence techniques to analyze microarray data. The aim of the system presented in this study is to provide innovative decision support techniques for classifying data from microarrays and for extracting knowledge about the classification process. The computational intelligence techniques used in this chapter follow the case-based reasoning paradigm to emulate the steps followed in expression analysis. This work presents a novel filtering technique based on statistical methods, a new clustering technique that uses ESOINN (Enhanced Self-Organizing Incremental Neuronal Network), and a knowledge extraction technique based on the RIPPER algorithm. The system presented within this chapter has been applied to classify CLL patients and extract knowledge about the classification process. The results obtained permit us to conclude that the system provides a notable reduction of the dimensionality of the data obtained from microarrays. Moreover, the classification process takes the detection of relevant and irrelevant probes into account, which is fundamental for subsequent classification and an extraction of knowledge tool with a graphical interface to explain the classification process, and has been much appreciated by the human experts. Finally, the philosophy of the CBR systems facilitates the resolution of new problems using past experiences, which is very appropriate regarding the classification of leukemia.

Keywords: Case-based Reasoning, HG U133, ESOINN, CLL leukemia classification, decision rules.

1 Introduction

The use of computational intelligence techniques has become fundamental in medicine, since there is a growing need of decision support tool that facilitate the monitoring of patients and the automatic processing of patient's data [1] [2] [3]. One of the fields in medicine requiring computational intelligence is the analysis of microarrays, and more specifically expression arrays, for the analysis of different sequences of

I. Bichindaritz et al. (Eds.): Computational Intelligence in Healthcare 4, SCI 309, pp. 289–312.
springerlink.com

oligonucleotides [1] [4]. The data obtained from microarrays are an important source of knowledge to prevent and detect cancer. The analysis of this information allows the detection of patterns that characterize certain diseases and, most importantly, the genes associated with these different diseases. Since the amount of data obtained from microarrays is huge and the time required to analyze the data is very high, it is necessary to obtained novel computational techniques that can provide automatic processing and artificial intelligence techniques that provide behaviours similar to the human ones.

An expression analysis basically consists of three stages: normalization and filtering; clustering and classification; and extraction of knowledge. These stages are carried out from the luminescence values found in the probes. Presently, the number of probes containing expression arrays has increased considerably to the extent that it has become necessary to use new methods and techniques to analyze the information more efficiently [5]. It is necessary to develop new techniques to analyze large volumes of data, extract the relevant information, and delete the information which has no relevance to the classification process. Moreover, the knowledge obtained during the classification process is of great importance for subsequent classifications. There are various artificial intelligence techniques such as artificial neural networks [6] [7], bayesian networks [8], and fuzzy logic [9] which have been applied to microarray analysis. While these techniques can be applied at various stages of expression analysis, the knowledge obtained cannot be incorporated into successive tests and included in subsequent analyses.

The system proposed in the context of this work focuses on the detection of carcinogenic patterns in the data from microarrays for patients, and is constructed from a CBR system that provides a classification technique based on previous experiences [11]. The system is an evolution of our previous work in the classification of leukemia patients [12], where a mixture of experts was used. The incorporation of the CBR paradigm to health care [13] [14] provides additional learning and adaptation capabilities. Moreover. The filtering and extraction of knowledge models have been improved and new techniques have been incorporated. The purpose of case-based reasoning (CBR) is to solve new problems by adapting solutions that have been used to solve similar problems in the past [10]. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential steps which are recalled every time that a problem needs to be solved: retrieve, reuse, revise and retain. Each of the steps of the CBR life cycle requires a model or method in order to perform its mission.

The approach presented in this work focuses on the classification of subtypes of leukemia, specifically, to detect patterns and extract subgroups within the CLL type of leukemia, and incorporates various techniques of computational intelligence at different stages of the reasoning cycle of a CBR system. In the retrieve phase, new pre-processing and filtering techniques are incorporated in order to select the probes with relevant information for classifying patients. This innovative method notably reduces the dimensionality of the data, which makes it possible to use techniques with greater computational complexity in later stages of the CBR cycle, which would otherwise be unviable. The reuse stage incorporates a classification technique based on ESOINN [15] neural networks, that proposes a novel method for generating clusters, and for identifying the nearest cluster for the final classification. An additional grouping technique known as PAM [16] (Partition around medoids), is also used, resulting in a more accurate classification, since the results suggested by the ESOINN network are compared to those obtained using the PAM technique. The revise phase initiates a RIPPER [43] algorithm for extracting knowledge about the classification process. Moreover, the revise stage includes a MDS (Multidimensional Scaling) technique [18] [19] [20] for presenting information in low dimensionality. Additionally, a human expert analyzes this information and evaluates the proposed classification as well as the validity of the rules generated. Finally, in the retain stage, if the human expert considers the proposed solution valid, the system stores the case information and the rules that have been obtained.

The chapter is structured as follows: the next section briefly introduces the problem that motivates this research. Section 3 presents the approach presented in this work and describes the novel strategies incorporated in the stages of the CBR cycle. Section 4 details the innovative computational intelligence techniques presented in this work. Section 5 describes a case study specifically developed to evaluate the CBR system presented within this study, consisting of a classification of CLL leukemia patients. Finally, Section 6 presents the results and conclusions obtained after testing the model.

2 Related Work

Microarray has become an essential tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease [21]. Microarray technology is based on a database of gene fragments called ESTs (Expressed Sequence Tags), which are used to measure target abundance using the scanned fluorescence intensities from tagged molecules hybridized to ESTs [22]. Specifically, the HG U133 plus 2.0 [5] are chips used for expression analysis. These chips analyze the expression level of over 47.000 transcripts and variants, including 38.500 well-characterized human genes. It is comprised of more than 54.000 probe sets and 1.300.000 distinct oligonucleotide features. The HG U133 plus 2.0 provides multiple, independent measurements for each transcript. The use of Multiple probes provides a complete data set with accurate, reliable, reproducible results from every experiment. Microarray technology is a critical element for genomic analysis and allows an in-depth study of molecular characterization of RNA expression, genomic changes, epigenetic modifications or protein/DNA unions.

Expression arrays [5] are a type of microarray~~s~~ that have been used in different approaches to identify the genes that characterize certain diseases [23] [24] [25]. In all cases, the data analysis process is essentially composed of three stages: normalization and filtering; clustering; and classification. The first step is critical to achieve both a good normalization of data and an initial filtering to reduce the dimensionality of the data set with which to work [26]. Since the problem at hand is working with high-dimensional arrays, it is important to have a good pre-processing technique that can facilitate automatic decision-making about the variables that will be vital for the classification process. In light of these decisions it will be possible to reduce the original dataset. Moreover, the choice of a clustering technique allows data to be grouped according to certain variables that dominate the behaviour of the group. After organizing into groups it is possible to extract knowledge and classify patients within the group which presents the most similarities.

Case-based reasoning [11] 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 [27]. CBR systems can be used to propose new solutions or evaluate solutions to avoid potential problems. The chapter of CBR in health care is discussed in [13] [14], where the advantages of this paradigm are remarked. The research in [28] suggests that analogical reasoning is particularly applicable to the biological domain, in part 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 [29] a mixture of experts for case-based reasoning (MOE4CBR) is proposed. It is a method that combines an ensemble of CBR classifiers with spectral clustering and logistic regression, but does not incorporate~~s~~ extraction of knowledge techniques and does not focus on dimensionality reduction. Some approaches such as [11] provide CBR solutions and knowledge extraction techniques, facilitating the comprehension of the classification process. This chapter presents a CBR solution which also incorporates new knowledge extraction techniques, but additionally focuses on the definition of innovative strategies for dimensionality reduction and clustering. The following section presents a detailed account of the CBR system proposed in this work.

3 CBR System as Paradigm for Classifying Microarray Data

This section presents the CBR system proposed in the context of this research and provides a classification technique based on previous experiences for data from microarrays. The CBR developed system imitates the behaviour of human experts in the laboratory and incorporates innovative knowledge discovery techniques. The system receives data from the analysis of chips and is responsible for classifying individuals based on evidence and existing data.

The purpose of CBR is to solve new problems by taking into account similar problems that were previously resolved in the past [10]. The primary concept when working with CBRs is the concept of case. A case can be defined as a past experience, and is composed of three elements: a problem description which describes the initial problem; a solution which provides the sequence of actions carried out in order to solve the problem; and the final stage which describes the state achieved once the solution was applied. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential steps which are recalled every time a problem needs to be solved: retrieve, reuse, revise and retain. Each step of the CBR life cycle requires a model or method in order to perform its mission.

In the CBR system proposed within this study, the retrieve phase filters variables, and recovers important variables from the cases to determine the most influential for the classification. Once the most important variables have been retrieved, the reuse phase begins adapting the solutions for the retrieved cases to obtain the clustering. Once this grouping is accomplished, the next step is knowledge extraction. The revise phase consists of an expert revision for the proposed solution, and finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases, consequently updating the cases memory.

A key element in a CBR system is a case, which can be defined as a past experience [10] and is composed of three elements: problem description, problem solution, and the final state obtained after applying the solution. A case in the system presented in this work contains information related to the patient, the rules, the proposed classification, and the probes marked as irrelevant or important. The case is defined by the following expression:

$$i_j = (id, S = (A_1, ..., A_n), C^p, C^r)$$

where $i_j \in I$ and $I = \{i_1, ..., i_s\}$ is the set of individuals/cases, A is the set of all the probes, A_i represents the probe i, C^p is the predicted class and C^r the actual class.

In addition to the cases memory, our system incorporates a memory of rules that contains the information extracted through the knowledge extraction techniques. The memory of rules is structured as follows: $R = \{r_1, ..., r_l\}$, with $r_i = (l_1 \wedge ... \wedge l_m) \rightarrow c_j$ where $l_s = (d_{ts}, o_s, \Re) / d_{ts} \in D_t, o_s \in O$, $S_{Irr} \subseteq A$

where R is the set of rules from the decision rules, l_s contains a set of discretized probes, an operator and a real value, D_t is the discretization value for the probe A_t, $O = \{=, \neq, >, <, \leq, \geq\}$ operator, and S_{Irr} is the set of probes marked as irrelevant, $c_j \in C^p$.

When a new case is classified, a new decision rules are generated in the revise stage. A set of rules are extracted which provide knowledge about the relevance of the probes in the clustering and classification process. Figure 1 shows a scheme of the techniques applied in the different stages of the CBR cycle. As seen in Figure 1, the important probes that allow the classification of patients are recovered in the Retrieve phase. The Retrieve phase is divided into 6 sub-phases: pre-processing through RMA, removal of irrelevant variables, uniform distribution, probes without meaningful cut-off points, and correlated variables. In the Reuse phase the patients are grouped by means of an ESOINN neural network. Then, the patients with no prior classification are assigned to a group using the nearest cluster. In the Revise phase the RIPPER [43] algorithm is applied for extracting knowledge about the most important probes for the classification, and the MDS technique [18] [19] [20] is used to make a representation in low dimensionality. Finally, in the Retain phase, the knowledge is updated. This knowledge includes the case classification, the decision rules obtained, and the information associated with the importance or irrelevance of certain probes extracted from the rules. Figure 1 shows the scheme of the CBR system proposed within this study.

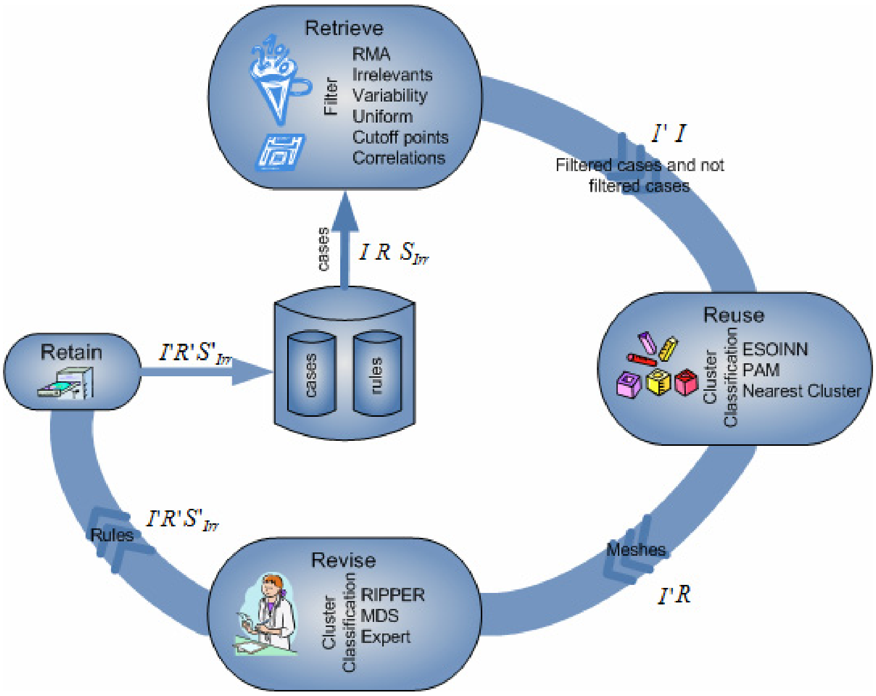


Fig. 1. Structure of the proposed CBR model

3.1 Retrieve

Traditionally, only the cases similar to the current problem are recovered, often because of performance, and then adapted. With regards to expression array, the number of cases is not a critical factor, rather the number of variables. For this reason, we have incorporated an innovative filtering strategy where variables are retrieved at this stage and then, depending on the identified variables, the rest of the stages of the CBR are carried out. The new strategy allows a notable reduction in the dimensionality of the data.

3.2 Reuse

In this phase the clustering of individuals is carried out, along with the classification of new individuals to one of the clusters. This chapter proposes a hybrid solution that takes into account a ESOINN neural network and the PAM method.

3.3 Revise

As shown in Figure 1, the revision is carried out by a human expert who determines if the group assigned by the system is correct. To facilitate the human expert task, the equivalence index and the error rate were calculated in the reuse stage. It is important for the medical human expert to understand the classification process performed in the two previous stages. In this sense, the system presented in this work provides a knowledge extraction method in the Revise phase. This method analyses the steps followed in the retrieve and reuse stages, and extracts knowledge which is then formalized in rules. In this way, the human expert can easily evaluate the classification and extract conclusions concerning the efficiency of the classification process. A RIPPER algorithm is used.

3.4 Retain

If the human expert identifies relevant information at the revise stage, the knowledge is acquired and the information obtained is stored. The information that is stored corresponds to the classifications considered correct, the decision rules generated that are considered relevant, and the probes marked as irrelevant. The information stored is divided into the cases memory I and the memory of rules R and S_{Irr}. Figure 1 shows the structure of the retain stage. Taking the revision of the expert into account, the system learns from the new experience and stores the information that the expert established as relevant. The stored information can include probes, classifications and rules.

4 Innovative Computational Intelligence Techniques for Dimensionality Reduction and Classification Improvement

This chapter details the innovative computational techniques included in the CBR phases of the system. The innovations consist of dimensionality reduction, classification improvements and extraction of knowledge technique. As the computational intelligence algorithms are included in the different phases of a CBR cycle, in this section we are going to present each of the novel algorithms as a part of the phases of the CBR cycle. Figure 1 details the steps followed in each of the stages of the CBR cycle. The structure of the CBR system proposed will now be explained in detail, presenting innovative techniques modelled in each of the stages of the CBR.

4.1 Filtering

This computational intelligence technique is carried out in the retrieve phase of the CBR cycle. The filtering phase is carried out on I together with the new case i_s. The filtering is only applied to those probes not associated with any of the rules. First, a pre-processing of the data is conducted using RMA. Then, the 5 filtering sub-phases are executed: removal of control probes, removal of erroneous probes, removal of low variability probes, removal of probes with a uniform distribution, and removal of correlated probes. These five sub-phases are outlined in the following paragraphs.

4.1.1 RMA

This phase begins once the laboratory experiment with microarrays has been completed. The researcher obtains various files that contain gross intensity values. Prior to analyzing the data, it is important to complete the pre-processing phase, which eliminates defective samples and standardizes the data. This phase is normally divided into 3 sub-phases: background correction, standardization, and summarization. There is currently a limited group of algorithms that investigators use for performing these steps. The most common are MAS5.0 [30] (Microarray Affymetrix Suite 5.0), PLIER [31] (Probe Logarithmic Intensity Error), and RMA) [32] (Robust Multi-array Average).

The RMA [32] algorithm is method for normalizing and summarizing probe-level intensity measurements. It analyzes the values for the PM (Perfect-Match): in the first step, a Background Correction is carried out to remove the noise from the averages of the PM; in the second step, the data is quantile normalized in order to compare data from different microarrays; finally, a summarization is made and the values for each probe-set are generated.

4.1.2 Irrelevant Probes

Once the control and the erroneous probes have been eliminated, the filtering process begins. The first step consists of eliminating the probes marked as irrelevant in previous executions of the CBR cycle. This way, all probes that can pass the filtering phase, but are prone to cause erroneous results during the reuse phase, are removed.

4.1.3 Variability

The second stage is to remove the probes that have low variability. This work is carried out according to the following steps:

1. Calculate the standard deviation for each of the probes j

$$\sigma_{.j} = +\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\overline{\mu}_{.j} - x_{ij}\right)^2} \qquad (1)$$

Where n is the total number of cases, $\overline{\mu}_{.j}$ is the average population for the variable j, and x_{ij} is the value of the probe j for the individual i.

2. Standardize the above values

$$z_i = \frac{\sigma_{.j} - \mu}{\sigma} \qquad (2)$$

Where $\mu = \frac{1}{n}\sum_{j=1}^{n}\sigma_{.j}$ and $\sigma_{.j} = +\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\overline{\mu}_{.j} - x_{ij}\right)^2}$ where $z_i \equiv N(0,1)$

3. Discard probes for which the value of z meets the following condition: $z < -1.0$. This will achieve the removal of about 16% of the probes if the variable follows a normal distribution.

4.1.4 Uniform Distribution

Finally, all remaining variables that follow a uniform distribution are eliminated. The variables that follow a uniform distribution will not allow the separation of individuals. Therefore, the variables that do not follow this distribution will be really useful variables in the classification of the cases. The contrast of assumptions is explained below, using the Kolmogorov-Smirnov [33] test as an example. H0: the data follow a uniform distribution; H1: the analyzed data do not follow a uniform distribution. Statistical contrast:

$$D = \max\{D^{+}, D^{-}\} \tag{3}$$

where $D^{+} = \max\limits_{1 \le i \le n}\left\{\frac{i}{n} - F_0(x_i)\right\}$ $D^{-} = \max\limits_{1 \le i \le n}\left\{F_0(x_i) - \frac{i-1}{n}\right\}$ with i as the pattern of entry, n the number of items and $F_0(x_i)$ the probability of observing values less than i with H_0 being true. The value of statistical contrast is compared to the next value:

$$D_\alpha = \frac{C_\alpha}{k(n)} \tag{4}$$

in the special case of uniform distribution $k(n) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$ and a level of significance $\alpha = 0.05$ $C_\alpha = 1.358$.

4.1.5 Cut-Off Points

This step removes the probes that, despite not following a uniform distribution, have no separation between elements, and do not allow the elements to be partitioned. The way to remove the probes is to detect changes in the densities of the data, and to select the final probes. The probes in which cut-offs or high densities are not detected are eliminated, as they do not provide useful information to the classification process. This will keep the probes that allow the separation of individuals. The detection of the separation intervals is performed by calculating the distance between adjacent individuals. Once the distance is calculated, it is possible to determine the potentially relevant values. The selection is carried out by applying confidence intervals for the values of these differences if the values follow a uniform distribution, or by selecting the values above a certain percentile if the values do not follow a normal distribution. This process is formalized as follows:

1. Let I' be the set of individuals with filtered probes together with the new individual, where $x_{.j}$ represents the probe j for all the individuals, and x_{ij} the individual i for the probe j
2. Select the probe $j = 1$, $x_{.j}$
3. Sort in increasing order values $x_{.j}$
4. Calculate the value for $x'_{ij} = x_{i+1j} - x_{ij}$
5. Determine if the variable x'_{ij} follows a uniform distribution by means of the Shapiro-Wilk test [34], otherwise go to step 10.

6. Calculate the value for $\bar{x}'_{.j}$
7. Establish the confidence interval for the variance, which is established as $\sigma'^{2}_{.j} \in \left[\frac{(n-1)\cdot S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)\cdot S^2}{\chi^2_{n-1,\alpha/2}}\right]$ with $\alpha = 0.05$ and $n = \# x'_{.j}$ and the number of elements for $x'_{.j}$, S is the sampling variance.
8. Establish the set of elements form x'_{ij} not belonging to the set $Q_j = \{x'_{ij} / x'_{ij} \notin I_{\sigma'_{.j}}\}$
9. Go to step 11.
10. Select those values up to the percentile P_α from every $x'_{.j}$ and establish the set $Q_j = \{x'_{ij} / x'_{ij} > P_\alpha\}$
11. Select the probe *j+1* in the case of more probes needing revision and go to step 2.
12. Create the new set of probes $I' = \bigcup x'_{.j} / \exists x'_{ij} \in Q_j / i > \# x' \cdot u \wedge i < \# x' - \# x' \cdot u$
13. Finalize and return the new set of individuals with the filtered probes I'

4.1.6 Correlations

At the last stage of the filtering process, correlated variables are eliminated so that only the independent variables remain. To this end, the linear correlation index of Pearson is calculated and the probes meeting the following condition are eliminated.

$$r_{x_i y_{.j}} > \alpha \tag{5}$$

given: $\alpha = 0.95$ $r_{x_i y_{.j}} = \frac{\sigma_{x_i x_{.j}}}{\sigma_{x_i}\sigma_{x_{.j}}}$, $\sigma_{x_i x_{.j}} = \frac{1}{N}\sum_{s=1}^{n}(\bar{\mu}_{.i} - x_{si})(\bar{u}_{.j} - x_{sj})$ where

$\sigma_{x_i x_{.j}}$ is the covariance between probes i and j.

4.2 Classification

There are several algorithms for clustering, but the most common are the hierarchical algorithms [35] and those based on partitioning [16]. Within the hierarchical algorithms the most common is the dendrogram [35]. The dendrograms are hierarchical methods that initially define conglomerates for each available case. At each stage the method joins the conglomerates with a smaller distance, and calculates the distance of the conglomerate with respect to the others. The new distances are updated in the distance matrix. The process finishes when there is only one conglomerate (agglomerative method) remaining.

Among the partition-based methods it is possible to find alternatives based on RNAs such as SOM [36] (Self-Organizing Map), GNG [37] (Growing Neural Gas) or

SOINN [38] (Self-Organizing Incremental Neuronal Network). Other alternatives are the methods based on heuristics, such as k-means [39] or PAM [16]. These two methods define a series of initial clusters which correspond to a new individual or to existing individuals, and are marked as the cluster representatives, while the remaining individuals are allocated to the nearest cluster. The problem with each of these methods is that they do not consider changes in the distribution of densities of individuals, and usually do not detect clusters with atypical forms, such as elongated clusters.

4.2.1 ESOINN Neural Network

Neural Networks based on GNG, allow detecting clusters with atypical forms, adjusting iteratively to the distribution of the individuals, and detecting low density zones. There are variants of the GNG, such as the GCS [40] (Growing Cell Structure) or SOINN [38] (Self-Organizing Incremental Neuronal Network). Unlike self-organizing maps based on meshes, Growing Grid or GCS do not set the number of neurons, or the degree of connectivity, but they do establish the dimensionality of each mesh. This complicates the separation phase between groups once the neurons are distributed evenly across the surface. The ESOINN neural network [15] (Enhanced Self-Organizing Incremental Neuronal Network) is a variation of the SOINN neural network [38], which allows the creation of a single layer, while ESOINN is able to incorporate both the distribution process along the surface and the separation between low density groups. The learning process of the network is distributed into two stages: the first stage of competition CHL [40] (Competitive Hebbian Learning) where the closest node to the input pattern is selected; and the second adaptation/growing stage similar to a GNG. The training phase and the various algorithms applied at every modified stage are outlined below:

1. Update the weights of neurons by following a process similar to the SOINN, but introducing a new definition for the learning rate in order to provide greater stability for the model. This learning rate has produced good results in other networks such as SOM [42].

$$\begin{gathered}\Delta W_{a_1} = n_1(M_{a_1})(\xi - W_{a_1}) \\ \Delta W_{a_i} = n_2(M_{a_1})(\xi - W_{a_i}) \text{ with } a_i \in N_{a_1}\end{gathered} \tag{6}$$

Where $n_1(x) = \frac{1}{\sqrt{x}}$, $n_2(x) = \frac{1}{\sqrt{2+x^2}}$, a_i is neuron i, ξ is the input pattern, M_{a_1} is the number of winnings of neuron a_i, N_{a_1} is the set of neighbours of a_i.

2. Delete the connections with higher age. The ages are standardized and those whose values are in the region of rejection with k>0 are removed. The assigned value of α is 0.05, therefore

$$z_i = \frac{e_i - \mu}{\sigma} \text{ , } z \equiv N(0,1) \text{ then } f(z) = \frac{1}{\sqrt{2\pi}} Exp\left[\frac{-z^2}{2}\right] \tag{7}$$

Where $P(z < k) = \alpha / 2$ → $P(z < k) = 0.975$ → $\Theta(z) = 0.975$ k=1.96
Therefore all z values that are greater than 1.96 are deleted

3. Once all input patterns have been introduced then a KS-Test [33] is carried out in order to determine if the density distribution for the neurons in each group follows a normal distribution. If so, the learning procedure is finished; otherwise the next pattern is processed. The value of α chosen is 0.05.

Once the cases have been distributed in the meshes, it is necessary to assign each of the meshes to a class according to the following procedure: Let I' be the set of individuals once the probes have been filtered and G^E the set of clusters created by means of the ESOINN neural network, defined as $G^E = \{g^E / g^E \subseteq I'\}$, where $g_i^E \cap g_j^E = \phi, \forall i \neq j$ with $g_i^E, g_j^E \in G^E$. Let C be the set of existing classes for the individuals where $c_j \in C$ is the class j in the set. We can say that the mesh i, g_i^E belongs to a class j, $g_i^E \in c_j$ and is represented as $g_{i_{c_j}}^E$ when

$$\max_{c_j \in C} \frac{\# I_{c_j} / g_i^E}{\# I_{c_j}} \cdot \frac{\# I'_{c_j}}{\# I'} \tag{8}$$

where $I'_{c_j} = \{s \in I / s \in c_j\}$ and I'_{c_j} / g_i^E is the set of individuals from c_j restricted to the group g_i^E.

The set of meshes belonging to the class is denoted as $G_{c_j}^E$ and is defined by the expression (9).

$$G_{c_j}^E = \bigcup_{g_{i_{c_j}}^E \in G^E} g_{i_{c_j}}^E \tag{9}$$

4.2.2 PAM

The PAM algorithm [16] is executed parallel to the clustering in order to facilitate a comparison of the results obtained. The classification made by both methods, PAM and ESOINN, generates an equivalence index between the two methods that determines the consistency of the reuse phase. The algorithm used for PAM is as follows:

1. Select the number of clusters depending on $\# C$.
2. The metric used for the distance is the same as the one used in the ESOINN network
3. Classify the patients taking all of the variables into account, without any filtering $G^P = \{g^P / g^P \subseteq I\}$ with $g_i^P \cap g_j^P = \phi, \forall i \neq j$
4. Once the groups G^P are created, an assignation is made following the procedure indicated in (8).

4.2.3 Equivalence Index

Once the individuals have been classified using both the PAM and the ESOINN neural networks, the equivalence index for both methods eq is calculated, and the error rate for the ESOINN network is determined as a function of the pre-classified cases. The equivalence index is defined as indicated in (10):

$$eq = \frac{\#\left\{i \in I' / i \in c_j^E \wedge i \in c_j^P\right\}}{\#I'} \tag{10}$$

Where c_j^E represents the set of meshes belonging to class j through the ESOINN network and c_j^P represents the set of individuals belonging to class j through the PAM algorithm.

4.2.4 Classification

Once the meshes are generated by the clustering process, previously unclassified individuals are now classified by selecting the nearest mesh. When the mesh has been selected, the case is assigned to the class of the mesh selected. The assignment is defined as (11).

$$i_s \in G_{k_{c_j}}^E \rightarrow i_u \in C_j \tag{11}$$

where i_u is the unclassified individual, i_s is the individual closest to the individual i_u calculated using the Euclidean distance.

As shown in Figure 1, the reuse stage receives the filtered and not-filtered data resulting from the retriever stage as inputs. The input is used for both the ESOINN neural network and the PAM technique. The ESOINN neural network generates a set of groups assigned to different classes. These groups are composed of meshes containing different elements together with the information of the previous classification. The PAM technique repeats the same project concurrently and generates the groups for each of the classes. The groups generated by PAM contain the individuals and their previous classification, but do not consider sub-groups. Finally, the equivalence index is calculated and the new patient is classified. The error rate for the ESOINN network is made through (8) to determine the class for each of the groups.

4.3 Knowledge Extraction

The knowledge extraction phase detects anomalous classifications, since it accounts for the existence of probes with irrelevant information, or those that were decisive for the misclassification. Sometimes, the existence of certain probes causes a classification of patients based on erroneous criteria, such as the distinction between men and women. Such a classification, without being wrong, is irrelevant to the problem, which is why the probes that can cause these classifications are analyzed at this stage. If the human expert notes that the probes contain irrelevant information, they are marked as irrelevant and not taken into account in the next iteration of the CBR cycle.

The extraction of knowledge that is presented to the human expert is carried out using the RIPPER [43]. There are other alternatives for the generation of decision rules which operate similar to the decision trees, including J48 [17] and PART [44]. These methods extract similar information to classify individuals according to decision rules. The results are similar for the different methods.

The general objective of extraction of knowledge techniques is to provide a human expert with information about the system-generated classification by means of a set of rules that support the decision-making process. It should be noted that knowledge extraction techniques are not intended to substitute the rationale and experience of a human expert during a diagnosis, rather to complement the process and serve as an additional methodology or guideline for common procedures in analysis.

The process is described in the following steps. Let I'be the set of individuals and s the number of probes once the filtering process has finished:

$$\begin{aligned} f_r : A_1 \times ... \times A_s \rightarrow A_1^{'} \times ... \times A_s^{'} \\ (a_1,...,a_s) \rightarrow f_r(a_1,...,a_s) = (a_1^{'},...,a_s^{'}) \end{aligned}$$

where $A_i^{'} \in [0,1]$ is the value of the term i using the function f_r and is obtained in the following way:

$$a_i^{'} = \frac{a_i - \min(A_i)}{\max(A_i) - \min(A_i)}$$

Finally the values are discretized by means of f_u in a series of predefined levels t.

$$\begin{aligned} f_u : A_1^{'} \times ... \times A_s^{'} \rightarrow \overbrace{D \times ... \times D}^{s} \\ (a_1^{'},...,a_s^{'}) \rightarrow f_u(a_1^{'},...,a_s^{'}) = (d_1^{'},...,d_s^{'}) \end{aligned}$$

where $D = \bigcup_{i \in \{0,...,t-1\}} i \cdot \frac{1}{t-1}$, we can say that $d_i^{'} = d_j$if applying the function f_u, with $d_j \in D$ if $d_i^{'} \in [d_j - 1/(2t),\ \ d_j + 1/(2t)]$.

Once the transformation is finished, the set of individuals is determined by the subset $I' \subseteq \overbrace{D \times ... \times D}^{s}$ of the data, and RIPPER is used to generate the rules that classify the individuals. The use of RIPPER, allows rules to be obtained for classifying an individual $i_k \in I'$ to the class c_j by means of rules similar to:

$$r_i = (l_1 \wedge ... \wedge l_m) \rightarrow c_j$$

where d_p is the value for the attribute p for the individual i_k. In this way the set is defined for rules R' that classify the individuals for each of the classes.

The input corresponds to the discretization of the values (if the reuse phase has been successful). Subsequently, knowledge extraction is applied through the RIPPER. Finally, the relevant information extracted is stored (probes inconsequential, important and results of the classification). At this stage a 3D representation with the information retrieved is displayed. The dimensionality is reduced by using MDS [18] [19] [20].

5 Case Study: Computational Intelligence Techniques for Classification of CLL Leukemia

Microarray analysis has made it possible to characterize the molecular mechanisms that cause several cancers. Regarding leukemia, microarray analysis has facilitated the identification of certain characteristic genes in the different variants of leukemia [24] [41] [45]. Cancer experts remark on the importance of the identification of the genes associated to each type of cancer in order to establish the most efficient treatments for the patients [46] [47]. The Cancer Institute in the city of Salamanca was interested in novel tools for decision support in the process of CLL (Chronic Lymphocytic Leukemia) patient classification. In this way, we focus on a concrete leukemia subtype, while our previous works were aimed at classifying patients into leukemia subtypes [12].

The Institute provided us with 91 samples of patient data and asked for a tool to provide decision support in the expression array analysis process and to incorporate innovative techniques to reduce the dimensionality of the data and identify the variables with a higher influence in the patient's classification. The samples corresponded to patients affected by chronic lymphocytic leukemia. CLL is a disease of lymphocytes that appear to be mature but are biologically immature. These B lymphocytes arise from a subset of CD5-B cells that appear to have a role in autoimmunity. The pathogenesis of chronic lymphocytic leukemia is likely a multistep process, initially involving a polyclonal expansion of CD5-B cells followed by the transformation of a single cell [48]. CLL is one of four main types of leukemia. About 15.110 new cases of CLL will be diagnosed in 2008. Approximately 90.179 people are currently living with CLL, more than the number of people living with any other type of leukemia. Most people with CLL are at least 50 years old [49]. CLL starts with a change to a single cell called a lymphocyte. Over time, the CLL cells multiply and replace normal lymphocytes in the marrow and lymph nodes. The high number of CLL cells in the marrow may crowd out normal blood-forming cells, and CLL cells are not able to fight off infection like normal lymphocytes do [49]. The aim of the tests performed in this study is to determine whether our system is able to classify new patients based on previously analyzed and stored cases.

6 Results and Conclusions

This chapter has presented a case-based reasoning system, that evolved from a previous work in leukemia patients classification [12], specifically designed to analyze data from microarrays, facilitating the grouping and classification of individuals. Moreover, the system provides an innovative method for exploring the classification process and extracting knowledge in the form of rules which help the human experts to understand the classification process and obtain conclusions about the relevance of the probes. The human experts in the laboratory have remarked on the advantages of using the system as a decision support system for CLL classification, and have especially noted the facility in acquiring knowledge and explanations.

Section 5 presented the case study considered in this report, which classified 91 CLL leukemia patients into groups. The aim of the case study was to identify the probes that allow classifying the CLL leukemia patients into subgroups. In an initial test, data from 91 patients, where previous classification was not taken into account, were used in the system. The pre-processing phase began with 54.675 probes and the RMA was applied to obtain the luminescence values for each of the probes and to homogenize the values from different chips. After the pre-processing phase, the filtering process was applied, notably reducing the probes to 541, without increasing the error rate.

Once the filtering was executed, it was still difficult to extract knowledge from the data. Figure 2 shows the 91 individuals in a bar graph, where the bars are divided in 541 probes with amplitude proportional to their value. The upper part of Figure 2 shows the classification obtained for each of the individuals, and the bottom of Figure 2 shows the parallel coordinates that represent 561 coordinates and 91 lines for each of the individuals.

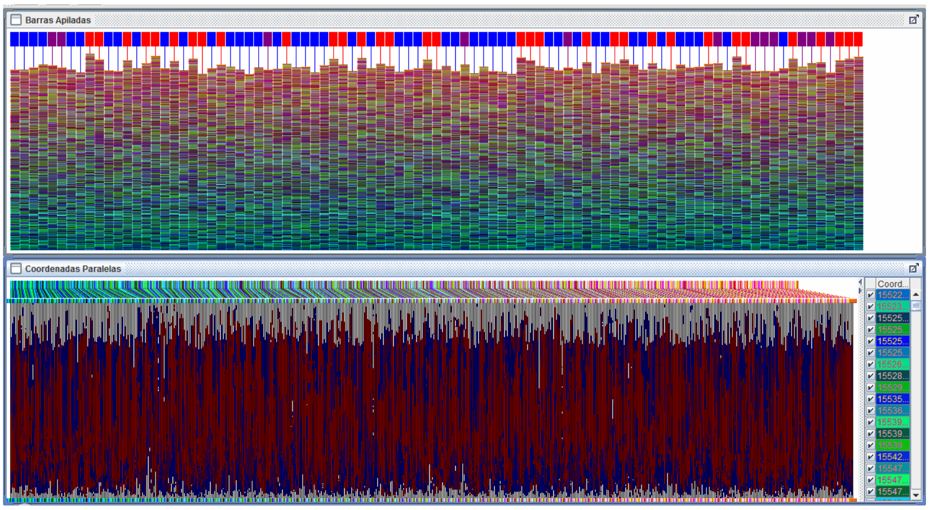


Fig. 2. Values for the probes obtained from the individuals.

The reuse phase begins when the probes have been filtered, and generates the meshes for the groups as well as the distribution of the individuals along the space. The mesh closest to the new case was then selected and the classification was made. To evaluate the proposed model, the system classified 90 individuals together with a new individual, and the results obtained were compared to the previous existing classifications. This process was repeated for each of the 91 individuals considered for the experiment and the results obtained demonstrate that 82 of the 91 individuals were successfully classified.

In the revise phase, the CBR system extracted the knowledge obtained during the classification process, as shown in Figure 3. Figure 3 presents the decision rules obtained in the revision phase that are applied to extract knowledge from the classification carried out in the previous phase.

(209083_at <= 0.25) and (1552280_at <= 0) => Class =C2 (10.0/1.0)
(231592_at >= 1) => Class=C2 (4.0/1.0)
(203213_at >= 0.75) => Class =C3 (29.0/0.0)
(1552619_a_at >= 0.5) => Class =C3 (2.0/0.0)
=> Class =C1 (46.0/1.0)

Fig. 3. Decision Rules obtained in the revision phase.

Figure 4 represents some graphics where the values of the retrieved probes are compared, and the information obtained from the retrieved probes is presented as decision rules. The values of the probes shown in Figure 4 are not the discretized ones used for the decision rules. At the top of Figure 4, both the real classification and the classification predicted by the system are presented by means of decision rules. If the colour matches, then there is a coincidence in the classification. As can be seen, there is an individual misclassified in the first of the classes identified in Figure 4, zero in the second class and two in the third class. At the bottom of Figure 4, it is possible to observe the parallel coordinates and the colours represent the class associated to the individual. As can be seen, it is possible to distinguish the probes associated to each of the classes. In this way, in the first of the coordinates can be seen how a group of individuals is separated from the rest.

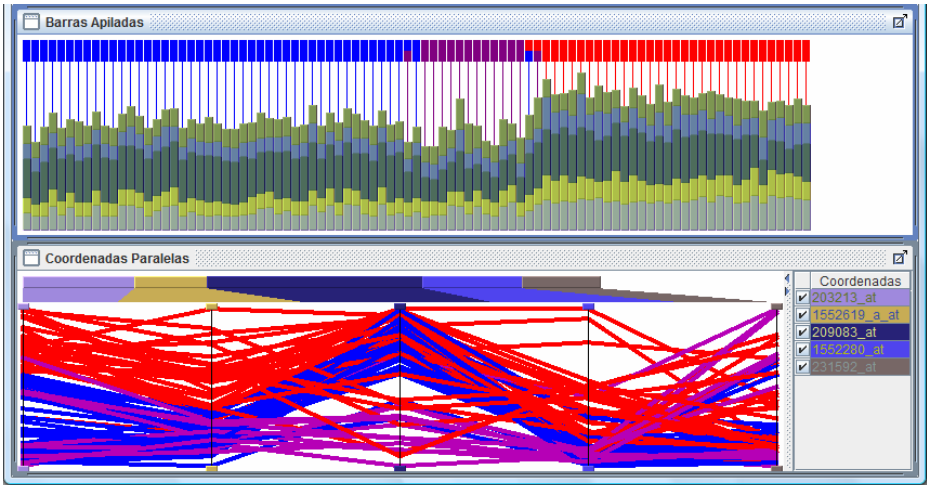
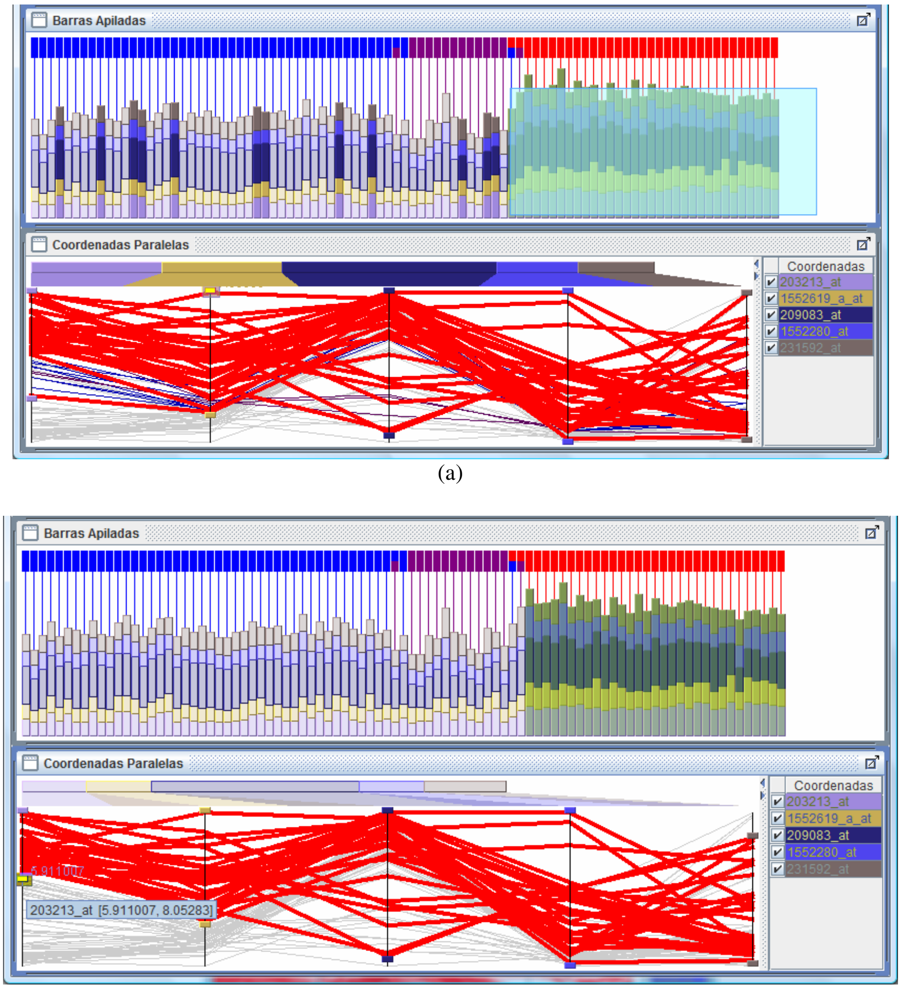


Fig. 4. Representation of the retrieved probes in terms of decision rules for the 91 individuals.

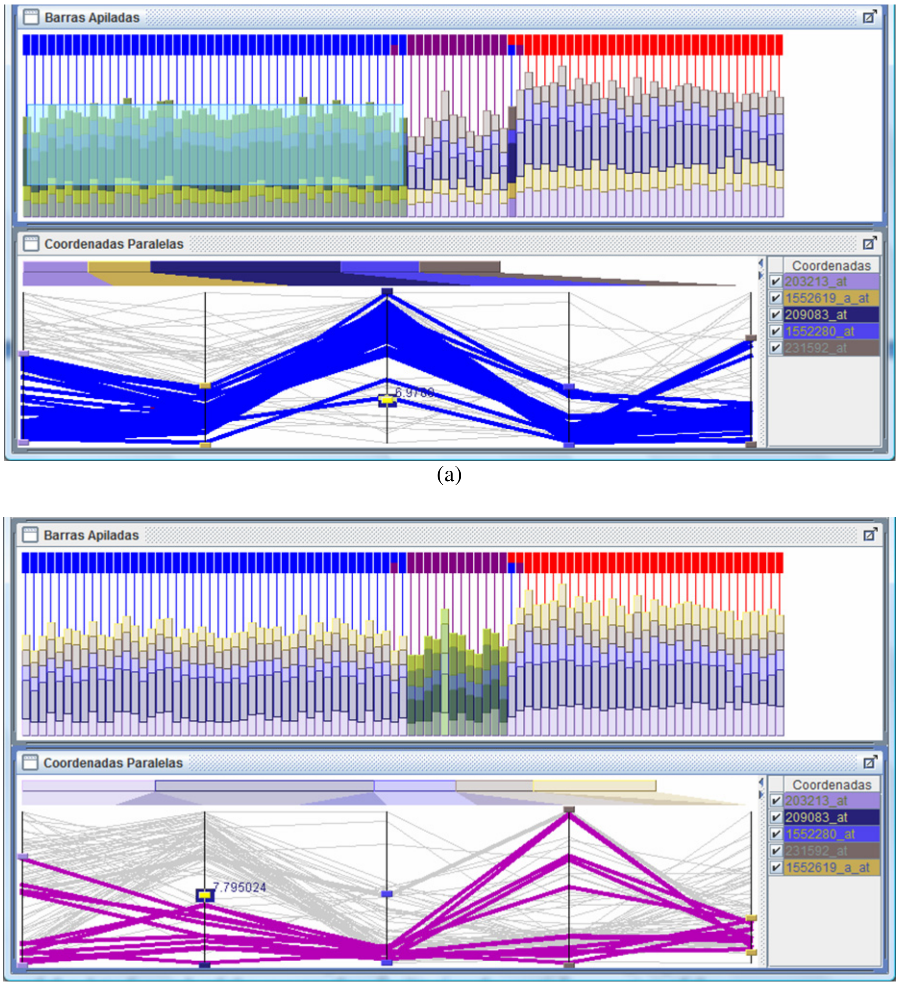
Figure 5 shows the classification of the individuals for the last class. Figure 5a shows the individuals classified to the class C3. In the classification of the individuals, the parallel coordinates establish the top and bottom margins for each of the probes, facilitating a graphical representation of the information contained by the rules. Once the margins are established, the individuals out of the ranks are shown as dimmed in the bars and the parallel coordinates. As can be seen in Figure 5a, when the individuals marked in red colour were selected, some individuals marked in blue (class C1) and marked in violet (class C2) were activated. Looking at the first of the coordinates, it is possible to observe a red line that corresponds to an individual with a low value and that is the responsible of the activation of the individuals of the C1 and C2 classes. Figure 5b shows a selection of the individuals estimated as members of the class C3. As can be seen in Figure 5b, the rest of the individuals remain dimmed, which allows a separation of the rest of the individuals. This is possible because of the information provided by the decision rules.



(a)

(b)

Fig. 5. Representation of the retrieved probes using the decision rules. (a) represents those individuals that are situated in the same rank of values than the individuals of the class C3. (b) shows the individuals situated in the same rank of values than the individuals estimated as members of the class C3.



(a)

(b)

Fig. 6. Representation of the probes for the individuals of the classes C1y C2.

Figure 6 represents the classification of individuals for the first class. As can be seen in Figure 6a, when the individuals of the class C1 were selected, only one of the individuals of the rest of the classes was activated. Figure 6b shows the results obtained when the margin of parallel coordinates was configured in order to avoid the activation of individuals of the other classes. As can be seen, only one individual of the class C2 was deactivated, which indicated that it was out of the margins, with a high value for the probe 15552280_at.

To obtain a visual representation of the patient's classification, we use the MDS [18] [19] [20], and the dimensionality of the data is reduced to three. Figures 7a and 8b represent the information once MDS has been applied and, as shown, the individuals of the different clusters are separated in the space. Figure 7 shows a representation

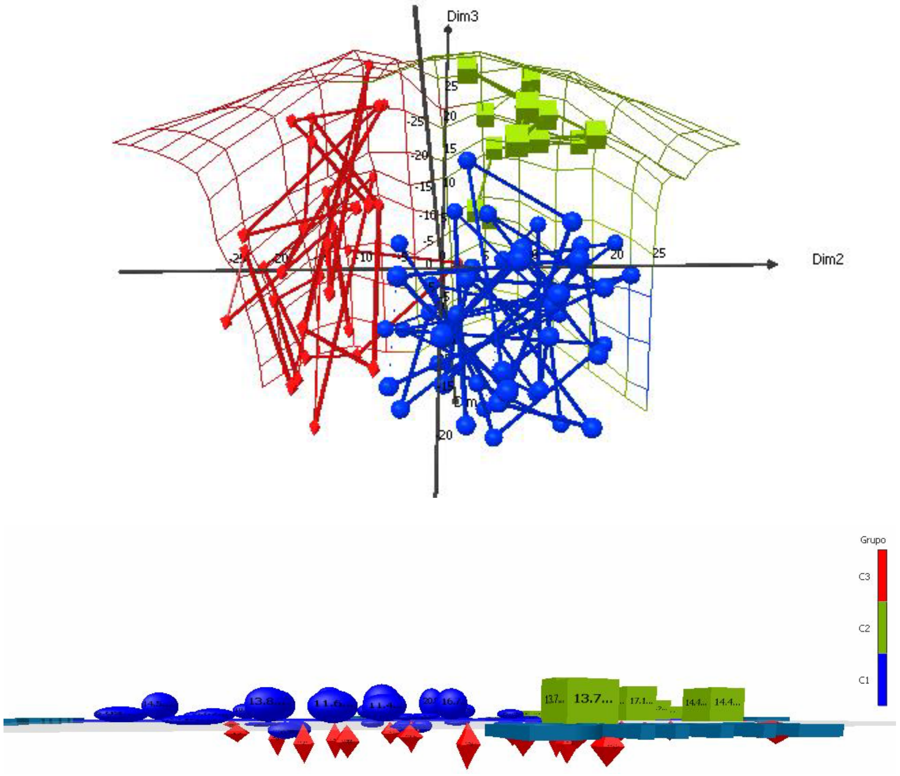


Fig. 7. Representation of low dimentionality probes with MDS

of the classification obtained for the individuals. Figure 7a represents the information obtained in a 3D format and, as can be seen, it is possible to identify three different classes clearly separated. Figure 7b shows a heatmap for the classification. As can be seen, class C1 contains negative values, while the classes C2 and C3 contain positive values.

In order to evaluate the global functioning of the system, we included an additional test. In this test the system contains 45 previously classified individuals, and aims to classify the remaining 46 cases using previous knowledge. Figure 8 presents the error rate identified for each of the interactions of the CBR system. As can be seen in Figure 8, the error rate is reduced after the initial iterations. The user of the CBR paradigm provides the ability for learning from previous experiences, which improves the performance of the classification process. In this sense, the classification provided by the system presented within this chapter improves the classification provided in our previous works 12 and provides a more detailed knowledge about the classification process.

The approach presented in this chapter is a specialized and novel system that integrates the steps of an expression analysis within the stages of a CBR cycle. The system is able to incorporate the knowledge acquired in previous classifications and use

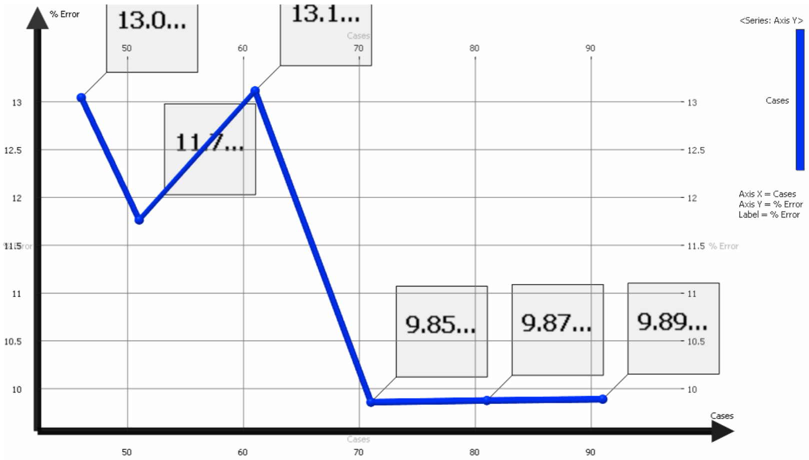


Fig. 8. Error rate for the classification process related to the iterations.

it to perform new classifications, providing a much appreciated decision support tool for doctors. As demonstrated, the proposed system reduces the dimensionality based on the filtering of genes with little variability and those that do not allow a separation of individuals due to the distribution of data. It also presents a clustering technique based on the neuronal network ESOINN, which is validated with a PAM technique. Finally, the system incorporates a technique for knowledge extraction and presents it to the human experts in a very intuitive format.

With the results obtained from empirical studies we can conclude that the CBR system presented in this study provides a tool that detects genes and probes, which are the most important factor for the detection of pathology, and facilitates a classification and reliable diagnosis, as shown by the results presented in this chapter. The system has been applied to classify CLL leukemia patients and allows the human expert to obtain information about the classification process and to identify the probes considered as important or irrelevant for further classifications. Taking into account these results, we can conclude that the incorporation of computational intelligence techniques in the expression analysis can facilitate the working day of the care personnel and provide a robust and reliable decision support tool for the prevention and detection of cancerous patterns.

Acknowledgments. Special thanks to the Institute of Cancer of Salamanca for the information and technology provided.

References

1. Tu, Y.J., Zhou, W., Piramuthu, S.: Identifying RFID-embedded objects in pervasive healthcare applications. Decision Support Systems 46(2), 586–593 (2008)
2. Chakraborty, D., Takahashi, H., Suganuma, T., Takeda, A., Kitagata, G., Hashimoto, K., Shiratori, N.: Context-aware remote healthcare support system based on overlay network. WSEAS Transactions on Computers 7(9), 1505–1514 (2008)

3. Lina, S., Chien, F.: Cluster analysis of genome-wide expression data for feature extraction. Expert Systems with Applications 36(2-2), 3327–3335 (2009)
4. Stadlera, Z.K., Come, S.E.: Review of gene-expression profiling and its clinical use in breast cancer. Critical Reviews in Oncology/Hematology 69(1), 1–11 (2009)
5. Affymetrix. GeneChip® Human Genome U133 Arrays, `http://www.affymetrix.com/support/technical/datasheets/hgu133arrays_datasheet.pdf`
6. Sawa, T., Ohno-Machado, L.: A neural network based similarity index for clustering DNA microarray data. Computers in Biology and Medicine 33(1), 1–15 (2003)
7. Bianchia, D., Calogero, R., Tirozzi, B.: Kohonen neural networks and genetic classification. Mathematical and Computer Modelling 45(1-2), 34–60 (2007)
8. Baladandayuthapani, V., Ray, S., Mallick, B.K.: Bayesian Methods for DNA Microarray Data Analysis. Handbook of Statistics 25(1), 713–742 (2005)
9. Avogadri, R., Valentini, G.: Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artificial Intelligence in Medicine 45(2-3), 173–183 (2009)
10. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993)
11. Riverola, F., Díaz, F., Corchado, J.M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence 22(3-4), 254–268 (2006)
12. Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of Experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine 46, 179–200 (2009)
13. Bichindaritz, I.: Role and Significance of Case-based Reasoning in the Health Sciences. KI 23(1), 12–17 (2009)
14. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: What's next? Artificial Intelligence in Medicine 36(2), 127–135 (2006)
15. Furao, S., Ogura, T., Hasegawa, O.: An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks 20(8), 893–903 (2007)
16. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
17. Saravanan, N., Cholairajana, S., Ramachandran, K.I.: Vibration-based fault diagnosis of spur bevel gear box using fuzzy technique. Expert Systems with Applications 36(2-2), 3119–3135 (2009)
18. Borg, I., Groenen, P.: Modern multidimensional scaling theory and applications. Springer, Heidelberg (1997)
19. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to nonmetric hypothesis. Psychometrika 29(1), 1–27 (1964)
20. Ture, M., Tokatli, F., Kurt, I.: Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications 36(2), 2017–2026 (2009)
21. Quackenbush, J.: Computational analysis of microarray data. Nature Review Genetics 2(6), 418–427 (2001)
22. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., Lockhart, D.H.: High density synthetic oligonucleotide arrays. Nature Genetics 21(1), 20–24 (1999)
23. Taniguchi, M., Guan, L.L., Basarab, J.A., Dodson, M.V., Moore, S.S.: Comparative analysis on gene expression profiles in cattle subcutaneous fat tissues. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 3(4), 251–256
24. Avogadri, R., Valentini, G.: Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artificial Intelligence in Medicine 45(2-3), 173–183 (2009)

25. Margalit, O., Somech, R., Amariglio, N., Rechav, G.: Microarray based gene expression profiling of hematologic malignancies: basic concepts and clinical applications. Blood Reviews 4(4), 223–234
26. Armstrong, N.J., Van de Wiel, M.A.: Microarray data analysis: From hypotheses to conclusions using gene expression data. Cellular Oncology 26(5-6), 279–290 (2004)
27. Jurisica, I., Glasgow, J.: Applications of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics 25(1), 85–95 (2004)
28. Aaronson, J.S., Juergen, H., Overton, G.C.: Knowledge Discovery in GENBANK. In: Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, pp. 3–11 (1993)
29. Arshadi, N., Jurisica, I.: Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains. IEEE Transactions on Knowledge and Data Engineering 17(8), 1127–1137 (2005)
30. Affymetrix. Statistical Algorithms Description Document, http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf
31. Affymetrix. Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf
32. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J.: Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. Biostatistics 4, 249–264 (2003)
33. Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition 34, 1625–1637 (2001)
34. Jurečkováa, J., Picek, J.: Shapiro–Wilk type test of normality under nuisance regression and scale. Computational Statistics & Data Analysis 51(10), 5184–5191 (2007)
35. Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425 (1987)
36. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, 59–69 (1982)
37. Fritzke, B.: A growing neural gas network learns topologies. In: Advances in Neural Information Processing Systems, vol. 7, pp. 625–632 (1995)
38. Shen, F.: An algorithm for incremental unsupervised learning and topology representation, Tokyo: Ph.D. thesis. Tokyo Institute of Technology (2006)
39. Redmond, S.J., Heneghan, C.: A method for initialising the K-means clustering algorithm using kd-trees. Pattern Recognition Letters 28(8), 965–973 (2007)
40. Martinetz, T.: Competitive Hebbian learning rule forms perfectly topology preserving maps. In: ICANN 1993: International Conference on Artificial Neural Networks, pp. 427–434 (1993)
41. Guinn, B., Gilkes, A.F., Woodward, E., Westwood, N.B., Muftia, G.J., Linchc, D., Burnett, A.K., Mills, K.I.: Microarray analysis of tumour antigen expression in presentation acute myeloid leukaemia. Biochemical and Biophysical Research Communication 333(5), 703–713 (2005)
42. Corchado, J.M., Bajo, J., De Paz, Y., De Paz, J.F.: Integrating Case Planning and RPTW Neuronal Networks to Construct an Intelligent Environment for Health Care. Expert Systems with Applications 36(3), 5844–5858 (2009)
43. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets, Machine Learning (1993)

44. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization, pp. 144–151. Morgan Kaufmann, San Francisco (1998)
45. Vogiatzis, D., Tsapatsoulis, N.: Active learning for microarray data. International Journal of Approximate Reasoning 47(1), 85–96 (2008)
46. Yang, T.Y.: Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern. Computational Statistics & Data Analysis 53(3), 756–765 (2009)
47. Leng, C.: Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data. Computational Biology and Chemisty 32(6), 417–425 (2008)
48. Foon, K.A., Rai, K.L., Gale, R.P.: Chronic lymphocytic leukemia: new insights into biology and therapy. Annals of Internal Medicine 113(7), 525–539 (1990)
49. Chronic Lymphocytic Leukemia (2008), The leukemia and lymphoma society, `http://www.leukemia-lymphoma.org/all_page.adp?item_id=7059`