Case-based reasoning as a decision support system for cancer diagnosis: A case study

Juan F. De Paz, Sara Rodríguez, Javier Bajo and Juan M. Corchado
Departamento de Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, España
{fcofds;srg;jbajope;corchado}@usal.es

Corresponding Author:

Name: Javier Bajo

Tlfn: +34 639771985

Fax: +34 923277101

Email: jbajope@upsa.es

Address: Compañía 5, 37002, Salamanca, Spain

Abstract

Microarray technology can measure the expression levels of thousands of genes in an experiment. This fact makes the use of computational methods in cancer research absolutely essential. One of the possible applications is in the use of Artificial Intelligence techniques. Several of these techniques have been used to analyze expression arrays, but there is a growing need for new and effective solutions. This paper presents a Case-based reasoning (CBR) system for automatic classification of leukemia patients from microarray data. The system incorporates novel algorithms for data mining that allow filtering, classification, and knowledge extraction. The system has been tested and the results obtained are presented in this paper.

Keywords

Case-based Reasoning, HG U133, ESOINN, leukemia classification, decision tree

1 Introduction

Microarray has become an essential tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease [11]. Currently, there are

several kinds of microarrays such as CGH arrays [25] (Comparative Genome Hybridization), and expression arrays [1]. Microarray technology is a critical element for genomic analysis and allows the study of molecular characterization of RNA expression, genomic changes, epigenetic modifications or protein/DNA unions. Microarray technology is based on a database of gene fragments called expressed sequence tags (ESTs), which are used to measure target abundance using the scanned fluorescence intensities from tagged molecules hybridized to ESTs [23]. Specifically, the HG U133 plus 2.0 [1] 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. The HG U133 plus 2.0 is comprised of more than 54.000 probe sets and 1.300.000 distinct oligonucleotide features. It 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.

Expression arrays have been used in different approaches to identify the genes that characterize certain diseases [16] [20] [19]. In all cases, the data analysis process is essentially composed of three stages: normalization and filtering; clustering; and classification. The first step is critical for achieving both a good normalization of data and an initial filtering to reduce the dimensionality of the data set with which to work [18]. Since the problem at hand is working with high-dimensional arrays, it is important to have a good pre-processing technique that facilitates 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. These stages can be automated and included in a CBR [8] (Case-Based Reasoning) system.

For some time now, we have been working on the identification of techniques to automate the reasoning cycle of several CBR systems applied to complex domains [8]. The objective of this work is to develop a CBR system that allows the identification of patients with various types of cancer. The model aims to improve the classification of cancer based on microarray data. The system proposed in this paper presents a new synthesis that brings several artificial intelligence subfields together (filter techniques, clustering, artificial neural networks and knowledge extraction). 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. Specifically, the system presented in this paper uses a model which takes advantage of two novel methods for analyzing microarray data: a technique for filtering data, and the ESOINN technique [24] (Enhanced Self-Organizing Incremental Neuronal Network) for clustering. The first method combines various filtering techniques to dramatically reduce the dimensionality of the data. The second allows clustering by incorporating both the distribution process of the entire surface of classification, and the separation between groups with low density among them.

The paper is structured as follows: the next section briefly introduces the problem that motivates this research, presents the proposed CBR-based model, and describes the novel strategies incorporated in the stages of the CBR cycle. Section 3 describes a case study specifically developed to evaluate the CBR system presented within this work, consisting of a classification of leukemia patients. Finally, Section 4 presents the results and conclusions obtained after testing the model.

2 CBR System for Classifying Microarray Data

Microarray analysis has allowed the characterization of the molecular mechanisms that cause several cancers. Focusing on leukemia, microarray analysis has facilitated the identification of certain characteristic genes in the different variants of leukemia [20] [5] [7]. 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 [29] [6]. The relationship between the chromosomal alterations and the prognosis of leukemia and lymphomas is well established. Recently, conventional array-based expression profiling has demonstrated that chromosomal alterations are associated with distinctive expression patterns. The system proposed in 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.

The CBR developed 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 adapting solutions that have been used to solve similar problems 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 state 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 of the steps of the CBR life cycle requires a model or method in order to perform its mission. The algorithms selected for the retrieval of cases should be able to search the case

base and select the problem and corresponding solution most similar to the new situation. In our case study, the algorithms conducted a filtering of variables, recovered important variables from the cases, and determined which where most influential in the classification process. Once the most important variables have been retrieved, the reuse phase begins, in which the solutions for the retrieved cases are adapted so that clustering may be obtained. 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.

Figure 1 shows a diagram of the techniques applied in the different stages of the CBR cycle. As can be 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 Robust Multi-array Average (RMA), removal of control probes, erroneous probes, low variability, uniform distribution, and correlated variables. In the Reuse phase the patients are grouped by means of an ESOINN neural network. Then, the patients without prior classification are assigned to a group. In the Revise phase the Classification and Regression Tree (CART) technique is applied for extracting knowledge about the most important probes for the classification. Finally, in the Retain phase, the knowledge is updated.

Next, the structure of the CBR system proposed within this paper is explained in detail, and the innovative techniques modelled in each of the stages of the CBR are presented.

2.1 Retrieve

Traditionally, only the cases similar to the current problem are recovered, often because of their performance, and then adapted. With expression arrays, the number of cases is not a critical factor, rather the number of variables. For this reason, we have incorporated an

innovative 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. Figure 2 describes the steps carried out during the filtering phase. 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.

2.1.1 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 Affymetrix Microarray Suite 5.0 (MAS5.0) [3], Probe Logarithmic Intensity Error (PLIER) [2], and RMA [22].

The RMA [22] algorithm is frequently used for pre-processing Affymetrix microarray data and consists of three steps: (i) Background Correction: probe-level data for each chip are background corrected independently using a probabilistic model; (ii) Quantile Normalization: the background corrected probe-level data on each chip are normalized to a common set of quantiles, derived from background corrected data from all chips, whereby the goal is to make the distribution of probe intensities the same for arrays; and (iii) Expression Calculation: performed separately for each probe set n.

2.1.2 Control

During this phase, all probes used for testing hybridization are eliminated. These probes have no relevance at the time that individuals are classified, as there are no more than a few control points which should contain the same values for all individuals. If they have different values, the case should be discarded. Therefore, the probes control will not be useful in grouping individuals.

2.1.3 Erroneous

On occasion, some of the measurements made during hybridization may be erroneous; not so with the control variables. In this case, the erroneous probes that were marked during the implementation of the RMA must be eliminated.

2.1.4 Variability

Once both the control and the erroneous probes have been eliminated, the filtering begins. The first 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

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

where N is the total number of cases, $\overline{\mu}_{\cdot j}$ 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

$$z_i = \frac{\sigma_{\cdot j} - \mu}{\sigma} \tag{2}$$

where $\mu = \frac{1}{N}\sum_{j=1}^{N}\sigma_{\cdot j}$ and $\sigma_{\cdot j} = +\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\bar{\mu}_{\cdot j} - x_{ij}\right)^2}$ 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.

2.1.5 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 [21] test as an example. H0: the data follow a uniform distribution; H1: the analyzed data do not follow a uniform distribution. Statistical contrast:

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

where $D^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F_0(x_i)\right\}$ $D^- = \max_{1\le i\le n}\left\{F_0(x_i) - \frac{i-1}{n}\right\}$ 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 $\alpha = 0.05$ $C_\alpha = 1.358$.

2.1.6 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.

2.2 Reuse

Once the probes are filtered and standardized, they produce a set of values with i = 1 ... t, j = 1 ... s where N is the total number of cases, s the number of end probes. The next step is to perform the clustering of individuals based on their proximity according to their probes. Given that it is not possible to establish a fixed number of groups in the initial stages, and taking into account the need of the system to initiate the clustering without a previous classification, a technique for unsupervised classification was used. There is a wide range of possibilities for these techniques. Some are artificial neural networks such as SOM [27] (Self-Organizing Map), GNG [4] (Growing Neural Gas) which is the union of the CHL [26] (Competitive Hebbian Learning) and NG [28] (Neural Gas) techniques, GCS [26] (Growing Cell Structure), Growing Grid or the SOINN [9] (Self-Organizing Incremental Neuronal Network), or the ART [31] (Adaptive Resonance Theory). Other methods, such as self-organized Kohonen maps, set the number of clusters in the initial phase of training when using the k-means learning method algorithm. For this reason these methods cannot be used for the problem at hand, since in this case the number of clusters is unknown. However, the

number of groups could be adjusted and the degree of waste compaction checked so that according to this value, the final number of groups could be set. Nevertheless, this solution would require too much computing time and it would be difficult to limit the number of groups to include. The self-organized maps have other variants of learning methods that base their behaviour on methods similar to the NG. They create a mesh that is adjusted automatically to a specific area. The greatest disadvantage, however, is that both the number of neurons that are distributed over the surface and the degree of proximity are set beforehand, resulting in the number remaining constant throughout the entire training process, thus complicating, to a certain extent, the adaptation of the mesh. Unlike the 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 it is distributed evenly across the surface. The ART networks can be considered as an alternative. They are unsupervised learning networks that facilitate the automatic detection of clusters and, in their latest versions, allow the incorporation of continuous patterns. The major disadvantage of these networks is the selection of the monitoring parameter [30] to determine the number of clusters. Another disadvantage is that the knowledge extraction is more complicated than in mesh-based networks, so learning is less evident.

After analyzing different techniques and the problems each one might present as applied to the situation at hand, we have decided to use a variation of neural network SOINN [9], called ESOINN [24]. Unlike the SOINN, ESOINN consists of a single layer, so it is not necessary to determine the manner in which the training of the first layer changes to the second. With a single layer, ESOINN is able to incorporate both the distribution process along the surface and the separation between low density groups. The operation and training of the network

presents many similarities with those used in GCS networks as far as distribution over the surface is concerned, but not as far as the dimensionality of the meshes. Nevertheless, it more closely resembles a merger between a CHL and a NG: it has characteristics of a network CHL in the initial phases of the algorithm, by which it could be understood as a phase of competition, while in a second phase, the network of nodes begins to expand just as with a NG network. The training phase and the various algorithms applied at every stage are detailed below:

1. Create an empty set of nodes A
2. Create an empty set of interconnections between nodes $C \subset AxA$
3. Insert two nodes in the A and assign random values to weights W.
4. Select a pattern p $\xi_p = (x_{p1},...,x_{ps})$ of the data set D' with dimension R^s where x_{pj} represents the intensity of luminescence in probe j of individual i.
5. Search nodes: a_1 node closest to the pattern of entry and a_2 the second node closest to the pattern of entry, with $a_i \in A$, using Euclidean distance as the measure distance.

$$a_1 = \arg\min_{a \in A} \left\| \xi_p - W_a \right\|, a_2 = \arg\min_{a \in A\text{-}\{a_1\}} \left\| \xi_p - W_a \right\| \quad (6)$$

Up to this point the same steps are taken without changing any of the CHL algorithm techniques. Next the part concerning the ESOINN network begins, and modifications are made in order to automatically adjust various parameters. W_a is the weights vector for neuron a.

6. If the distance $a_1 > T_1$ or $a_2 > T_2$ add a new node in that position and continue with the step

$$T_i = \begin{cases} \max_{j \in N_i} \left\| W_i - W_j \right\| & N_i \neq \phi \\ \max_{j \in A-\{i\}} \left\| W_i - W_j \right\| & N_i = \phi \end{cases} \quad (7)$$

with $N_i \subseteq A$ being the set of neighbouring nodes of i

7. Increase the age of nodes connected with a_1 in one unit,

$$e_i(t+1) = e_i(t) + 1 \tag{8}$$

Where e_i represents the node i connected to a_1

8. If necessary, establish a new connection or delete it between a_1 and a_2 according to the following constraints:

- If either of the nodes is not associated with any subclass or both are in the same, then create a new connection between them and assign an age value of zero.
- If both nodes are in different subclasses A, B, then calculate the greater density for the neurons in each subclass A_{max} B_{max} so that if any of the following conditions are true, the subclasses are joined and a new connection is created; otherwise delete any connection that might exist.

$$\begin{aligned} \min(a_1) &> \alpha_A A_{\max} \\ \min(a_1) &> \alpha_B B_{\max} \end{aligned} \tag{9}$$

$$\text{With } \alpha_a = \begin{cases} 0.0 & 2.0 * mean_A \geq A_{\max} \\ 0.5 & 3.0 mean_A \geq A_{\max} > 2.0 * mean_A \\ 1.0 & A_{\max} > 3.0 * mean_A \end{cases} \quad mean_A = \frac{1}{N_A} \sum_{i \in A} h_i$$

This ensures that connections with nodes with a higher age neighbouring a_1 be subsequently deleted.

9. Update the density of a1 depending on the distance to the neighbour nodes

$$h_i = \frac{1}{N} \sum_{j=1}^{n} \sum_{k=1}^{\lambda} p_{ik} \tag{10}$$

where N is the number of $\sum_{k=1}^{\lambda} p_{ik} \neq 0$, n the number of periods, λ the number of patterns per period $\lambda = 50 + \sqrt{\#D}$

$$p_{ik} = \begin{cases} \dfrac{1}{\left(1 + \dfrac{1}{m}\sum_{j=1}^{m} \left\| W_i - W_j \right\|\right)^2} & i\ winner \\ 0 & eoc \end{cases} \tag{11}$$

where m is the number of neighbours to node i.

10. Increase the number of winning times winner for the neuron

$$M_{a_1}(t+1) = M_{a_1}(t) + 1 \tag{12}$$

11. 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 [13].

$$\begin{aligned} \Delta W_{a_1} &= n_1(M_{a_1})(\xi - W_{a_1}) \\ \Delta W_{a_i} &= n_2(M_{a_1})(\xi - W_{a_i}) \text{ with } a_i \in N_{a_1} \end{aligned} \tag{13}$$

Given $n_1(x) = \frac{1}{\sqrt{x}}$, $n_2(x) = \frac{1}{\sqrt{2 + x^2}}$ $a_i \in A$ is the neuron i, and N_{a_1} is set of neighbours a_i.

12. 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

$$z_i = \frac{e_i - \mu}{\sigma} \ , \ z \equiv N(0,1) \text{ then } f(z) = \frac{1}{\sqrt{2\pi}} Exp\left[\frac{-z^2}{2}\right] \tag{14}$$

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

13. If the number of iterations is a multiple of λ then carry out the following steps:
    1. Update subclass to which each neuron belongs bearing in mind the highest local density of each neuron with its neighbours.
    2. Delete nodes that meet any of the following conditions
        1. $N_a = 2$, $a \in A$ if $h_a < c_1 \sum_{j=1}^{\#A} h_j / \#A$ Where #A is the number of nodes of A, N_a are the neighbour nodes for a, c_1=0.001
        2. $N_a = 1$ con $a \in A$, si $h_a < c_2 \sum_{j=1}^{\#A} h_j / \#A$ Where #A is the number of nodes of A, c_2=1.0
        3. $a \in A$ and $N_a = 0$ delete a
14. The clustering of elements is carried out bearing in mind the connections among the neurons.
15. If all input patterns have been passed then a KS-Test [21] is carried out in order to determine if the density distribution for the neurons in each group follows a normal distribution. If so then the learning procedure is finished; otherwise the next pattern is processed. The value of α chosen is 0.05.

Once the meshes have been generated, previously unclassified individuals are classified by selecting the nearest mesh. Once the mesh has been selected, the case is assigned to the group with a high quantity of recovered elements. The allocation process is based on priorities. The individuals with the highest proportion are high priority level.

2.3 Revise and Retain

As shown in Figure 1, the revision is carried out by an expert who determines the correction with the group assigned by the system. If the assignation is considered correct, then the retrieve and reuse phases are carried out again so that the system can be ready for the next classification. If a classification is considered as incorrect or presents certain doubts, the case is not included into the memory of cases until the medical diagnosis is certain. It is important for the medical human expert to understand the classification process made in the two previous stages. For this reason, the CBR system proposed in this work incorporates 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 formalized in the set of rules. In this way, the human expert can easily evaluate the classification and extract conclusions on the efficiency of the classification process.

In the Revise stage, the data are initially discretized in five levels [0, 0.25, 0.5, 0.75, 1], and then the extraction of knowledge using the CART [14] algorithm is carried out. Finally the expert assigns the individual to the final group. The CART algorithm is a non parametric test that allows extracting rules that explain the classification carried out in the previous steps. There are other techniques to generate the decision trees, such as the methods based on Induction Decision Trees (ID3) [12], although currently CART is the most commonly used. This method allows rules to be generated and the most important variables to be extracted so that patients can be classified with a high degree of performance. The general objective of knowledge extraction techniques is to provide a human expert with information about the system-generated classification by means of a set of rules that are provided to 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.

Nevertheless, the system provides an automatic temporal revision for considering the retrieved cases. In the Retain stage, the system calculates the percentage of cases that have already been accurately classified among those retrieved for the current problem. If the percentage of a class is greater than the threshold, the system determines that the case has been successfully classified, and both the case and the knowledge obtained are stored in the memory of cases. This decision has to be confirmed by the human expert.

3 Case Study: Classification of Leukemia Patients

The Cancer Institute in the city of Salamanca was interested in novel tools for decision support in the process of leukemia patient classification. The Institute provided us with patient data and asked for a tool to automate certain tedious tasks in the expression array analysis process and incorporate innovative techniques to reduce the dimensionality of the data and identify the variables with a higher influence in the patient's classification. In the case study presented within this research, 212 samples were made available from analyses performed on patients either through punctures in the marrow or from blood samples. The samples corresponded to patients affected by five different types of leukemia: ALL (Acute Lymphocytic Leukemia), AML (Acute Myeloid Leukemia), CLL (Chronic Lymphocytic Leukemia), CML (Chronis Myeloid Leukemia) and MDS (Myelodysplastic Syndromes). The aim of the tests performed was to determine whether the system is able to classify new patients based on the cases previously analyzed and stored.

Figure 1 represents the bio-inspired model intended to resolve the problem of leukemia patient classification. The proposed model follows the procedures that are performed in medical centres. As can be seen in Figure 1, there is a previous phase which is external to the

model. This phase consists of a set of tests which have been carried out by laboratory personnel and allow us to obtain data from the chips. When a new sample is received, it is introduced into the chip. The chips are hybridized and explored by a scanner, allowing us to obtain information on the marking of several genes based on luminescence values. At that point, the CBR-based model starts to process the data obtained from the microarrays.

Figure 1 illustrates the phases of the CBR cycle, as well as the inputs and the outputs of each of the phases. The Retrieve phase uses a matrix D containing the data d_i of the individuals from the memory of cases and the new case. This phase recovers the relevant probes for patient classification. The total number of probes selected after pre-processing the data is represented as n. The system applies filtering techniques to remove control variables, erroneous variables, low variability variables, uniform distribution and correlated variables. The final number of probes remaining after the filtering process is represented as s, and D' represents the new reduced data matrix containing the luminescence intensities corresponding to individuals for the probes $d_i^{'}$.

As can be seen in Figure 1, the results obtained in the Retrieve phase are used as inputs for the Reuse phase. In this phase, patient classification is made by selecting the individuals that are most similar to the new individual, and by selecting the nearest cluster. The ESOINN neural network [24] uses these data to obtain clusters. The results of the Reuse phase consist of a set of meshes for the individuals representing the groups g_i, so that an unclassified user $d_{t+1}^{'}$ is assigned to the most numerous group in the closest mesh g_i.

Finally, during the Revise and Retain phases, the system uses the CART algorithm to obtain the decision rules used in the classification process, thus providing a set of rules f_i indicating the membership to each of the groups g_i. Each one of the rules contains a decision value v_i

and the probe z_i. The human expert determines if the classification process has provided successful results and indicates the way to proceed in the Retain phase.

4 Results and Conclusions

The CBR system presented in this work focused on identifying the important variables for each of the variants of blood cancer so that patients can be classified according to these variables. The model combines techniques for reducing the dimensionality of the original data set and a novel clustering method for classifying patients. The system works in a way similar to how human specialists operate in the laboratory, but is able to work with great amounts of data and make decisions automatically, thus significantly reducing both the time required to make a prediction, and the rate of human error due to confusion.

When conducting a study of leukemia based on data from microarrays, the process of filtering data takes on special importance. In the experiments reported in this paper, we worked with a database of bone marrow cases from 212 adult patients with five types of leukemia. The retrieve stage of the proposed CBR system presents a novel technique to reduce the dimensionality of the data. The initial number of probes in the experiment was 54.000, and the configuration obtained after the filtering process reduced this number to 785 probes, considered as really meaningful for the classification process. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer. The error rates have remained fairly low especially for cases where the number of patients was high.

We configured different settings in order to evaluate the global behaviour of the CBR system depending on the filtering strategy, which can be more or less restrictive. Table 1 shows the different values tested in each of the phases of the filtering process, as well as the

number of probes and individuals misclassified using the ESOINN neuronal network. The results presented in Table 1 show how 785 probes allow the classification of individuals with an error rate (30) similar to alternative configurations that take a greater number of probes into account.

Table 2 shows the total number of patients from each group and the number of misclassifications. As can be seen, groups with fewer patients are those with a greater error rate. The results shown in Table 2 are those obtained for the classification provided by the ESOINN neural network. The network learnt from all the patient data, after which each of the misclassified individuals was selected and classified to a group according to the classification of the other individuals. As CBR systems need initial knowledge to work in an efficient manner, 30 previously classified individuals were initially included in the memory of cases. A systematic sampling was applied in the selection of the 30 individuals from the 212 existing ones. No statistical technique was applied for the selection of the initial size of the sample since all the individuals are ultimately introduced into the system. The rest of the individuals were classified using the automated system proposed within this work. Finally, the 30 initial individuals were marked as unclassified and also assigned to the groups by the CBR system. Figure 3, shows the evolution of the error rate in the system.

The final classification was compared with the data obtained using a dendrogram [17] and PAM [15] (Partitioning Around Medoids). The proportion of errors in every group was calculated and the Kurskal-Wallis [32] test was applied to determine if the median of these proportions was equal. The results in table 3 show that after, applying statistical tests, the median of the proportions are different in each case. In addition, the value is lower than that obtained for other techniques. The asterisk * represents the values that are considered

different and (-) indicates that the value of the median of the technique in the column is better than the median of the technique in the row.

Figure 4 represents the distances matrix for leukemia patients, which takes into account the classification obtained in the Reuse phase and the classification provided by the medical staff. As seen in Figure 4, the distances between individuals of the same class are lower than those between individuals of different classes. Figure 4 shows two classes ALL and CLL where the patients can be clearly identified and distinguished from the rest of the patients. However, it is difficult to classify patients belonging to the AML, CML and MDS classes.

Once it can be verified that the retrieved probes allow classifying the patients in a way similar to the original classification, we can conclude that the retrieve phase works satisfactorily. The knowledge extraction is then carried out taking the selected probes into consideration. The algorithm used was CART [28], and the results obtained are shown in Figure 5. Figure 5 shows the probe and the condition, the total number of elements, the number of misclassified elements, and finally, the probability of assigning each of elements from the node to each of the groups, sorted as (ALL, AML, CLL, CML, MDS). The leaf nodes are identified by an asterisk *.

Figure 6 presents the decision tree generated through the information contained in Figure 5. The leaf nodes contain the classification for the individuals sorted by type of leukemia (ALL, AML, CLL, CML, MDS), and the intermediate nodes contain the conditions and values for the probes. When a condition is satisfied, the left branch is chosen. The most important probes and their relevance in the classification of patients are extracted by this algorithm. Figure 7 represents the first three probes retrieved with CART. Figure 7 shows that the CLL patients can be easily separated from the rest of the patients.

In the next step, the most significant probes of the CLL leukemias were extracted. These probes are shown in Figure 8. Figures 9a, 9b, 9c and 9d show the box plots for the four most significant probes for CLL patients. These probes are the same as those shown in Figure 7. As can be observed, the values of these probes are very different from the values for the rest of the groups. The model we propose resolves this discrepancy by using a technique that analyzes the available data in order to detect the genes of importance for the classification of diseases. As demonstrated, the proposed system reduces the dimensionality by filtering 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 in neuronal networks. The results obtained from empirical studies provide a tool that allows both the detection of genes and the most important variables for detecting pathology, and the facilitation of a classification and reliable diagnosis, as shown by the results presented in this paper.

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

5 References

[1] Affymetrix. GeneChip® Human Genome U133 Arrays http://www.affymetrix.com/support/technical/datasheets/hgu133arrays_datasheet.pdf

[2] Affymetrix. Guide to Probe Logarithmic Intensity Error (PLIER) Estimation http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf

[3] Affymetrix. Statistical Algorithms Description Document, http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

[4] B. Fritzke, A growing neural gas network learns topologies, *Advances in Neural Information Processing Systems 7*, (1995) 625-632.

[5] B. Guinna, A.F. Gilkesb, E. Woodwardb, N.B. Westwooda, G.J. Muftia, D. Linchc, A.K. Burnettb and K.I. Millsb Microarray analysis of tumour antigen expression in presentation acute myeloid leukaemia, *Biochemical and Biophysical Research Communications*, 333 (5) (2205) 703-713.

[6] C. Leng, Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data, *Computational Biology and Chemistry*, In Press.

[7] D. Vogiatzis and N. Tsapatsoulis, Active learning for microarray data, *International Journal of Approximate Reasoning*, 47 (1) (2008) 85-96.

[8] F. Riverola, F. Díaz and J. M. Corchado, Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets, *Computational Intelligence*, 22 (2006), 254-268.

[9] F. Shen, *An algorithm for incremental unsupervised learning and topology representation*, Tokyo: Ph.D. thesis. Tokyo Institute of Technology, 2006

[10] J. Kolodner, Case-Based Reasoning, *Morgan Kaufmann* 1993.

[11] J. Quackenbush, Computational analysis of microarray data, *Nature Review Genetics*, 2 (6) (2001) 418-427.

[12] J. Quinlan, Discovering rules by induction from large collections of examples, *Expert systems in the micro electronic age*, (1979) 168-201.

[13] J.M. Corchado, J. Bajo, Y. De Paz, J.F. De Paz Integrating Case Planning and RPTW Neuronal Networks to Construct an Intelligent Environment for Health Care, *Expert Systems with Applications*, 36 (2009) 5844–5858.

[14] L. Breiman, J. Friedman, A. Olshen and C. Stone, Classification and regression trees, *Wadsworth International Group*, 1984.

[15] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, *Wiley, New York,* 1990.

[16] M. Taniguchia, L.L. Guana, J.A. Basarabb, M.V. Dodsonc and S.S. Moorea Comparative analysis on gene expression profiles in cattle subcutaneous fat tissues, *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics.* 3 (4) 251-256.

[17] N. Saitou, M. Nie, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol. Biol,* 4 (1987) 406-425.

[18] N.J. Armstrong and M.A. Van de Wiel, Microarray data analysis: From hypotheses to conclusions using gene expression data, *Cellular Oncology*, 26 (5-6) (2004) 279-290.

[19] O. Margalit, R. Somech, N. Amariglio and G. Rechav, Microarray based gene expression profiling of hematologic malignancies: basic concepts and clinical applications, *Blood Reviews* 4 (4) (2005) 223-234.

[20] R. Avogadri and G. Valentini, The Corresponding Author and Giorgio Valentini Fuzzy ensemble clustering based on random projections for DNA microarray data analysis, *Artificial Intelligence in Medicine*, In Press

[21] R. Brunelli, Histogram Analysis for Image Retrieval. *Pattern Recognition*, 34, (2001) 1625-1637.

[22] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf and T.P. Speed, Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data, *Biostatistics*, 4 (2003) 249-264.

[23] R.J. Lipshutz, S.P.A. Fodor, T.R. Gingeras and D.H. Lockhart, High density synthetic oligonucleotide arrays, *Nature Genetics*, 21 (1999) 20-24.

[24] S. Furao, T. Ogura and O. Hasegawa, An enhanced self-organizing incremental neural network for online unsupervised learning, *Neural Networks*, 20 (2007) 893-903.

[25] Shinawi1 M. and Cheung S.W. The array CGH and its clinical applications, *Drug Discovery Today,* 13 (17-18) (2008) 760-770.

[26] T. Martinetz, Competitive Hebbian learning rule forms perfectly topology preserving maps, *ICANN'93: International Conference on Artificial Neural Networks,* (1993) 427-434.

[27] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, (1982) 59-69.

[28] T. Martinetz and K. Schulten, A neural-gas network learns topologies, *Artificial Neural Networks*, (1991) 397-402.

[29] T.Y. Yang. Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern, *Computational Statistics & Data Analysis.* In Press.

[30] A. Akhbardeh, Nikhil, P.E. Koskinenb and O. Yli-Harjaa Towards the experimental evaluation of novel supervised fuzzy adaptive resonance theory for pattern classification, *Pattern Recognition Letters*, 29 (8) 2008 1082-1093.

[31] G.A. Carpenter and Grossberg, S., The ART of adaptive pattern recognition by a self-organizing neural network. *IEEE Trans. Computer*, (1987) 77–88.

[32] W. Kruskal and W. Wallis, Use of ranks in one-criterion variance analysis, *Journal of American Statistics Association* (1952).

Table 1. Plans of the filtering phase and plan of greater efficiency

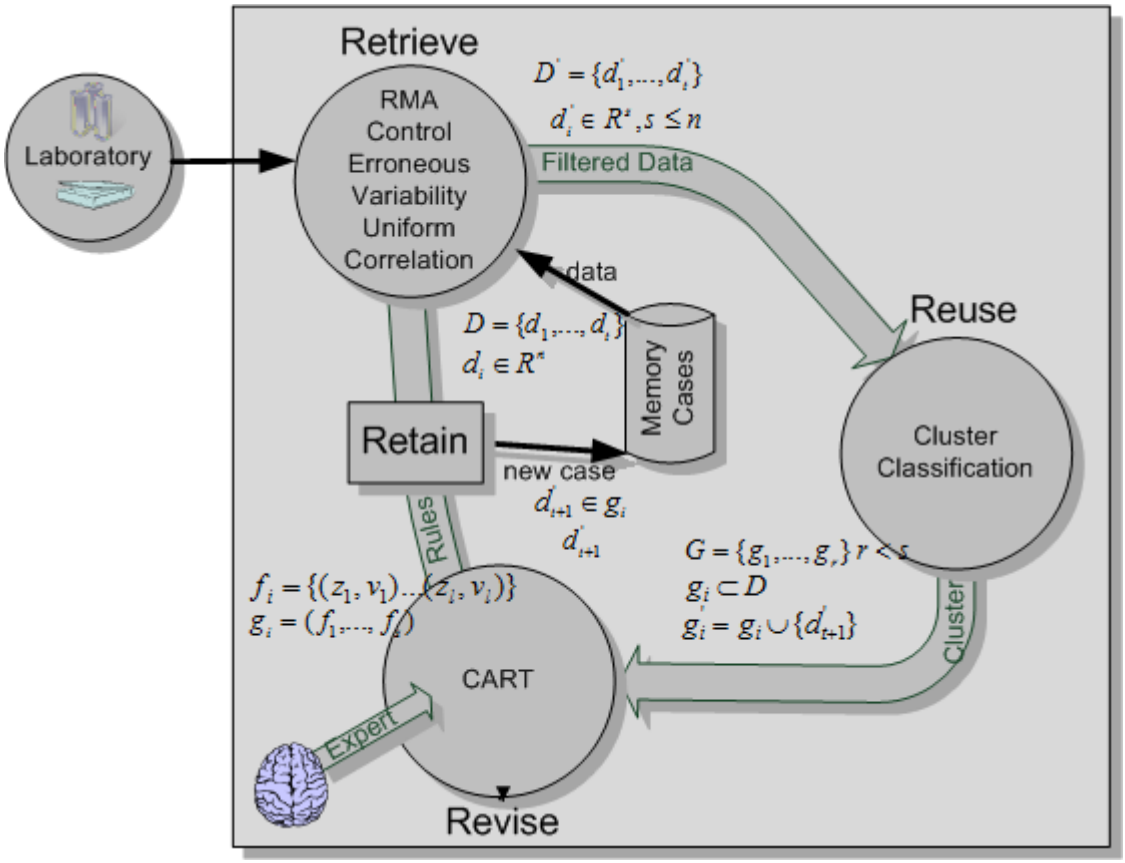
Variability (z)	Uniform (α)	Correlación (α)	Probes	Errors
-1.0	0.25	0.95	2675	27
-1.0	0.15	0.90	1341	28
-1.0	0.15	0.95	1373	28
-0.5	0.15	0.90	1263	30
-0.5	0.15	0.95	1340	29
-1.0	**0.1**	**0.95**	**785**	**30**
-1.0	0.05	0.90	353	47
-1.0	0.05	0.95	357	45
-0.5	0.05	0.9	332	67
-0.5	0.05	0.95	337	67
-1.0	0.01	0.95	54	83

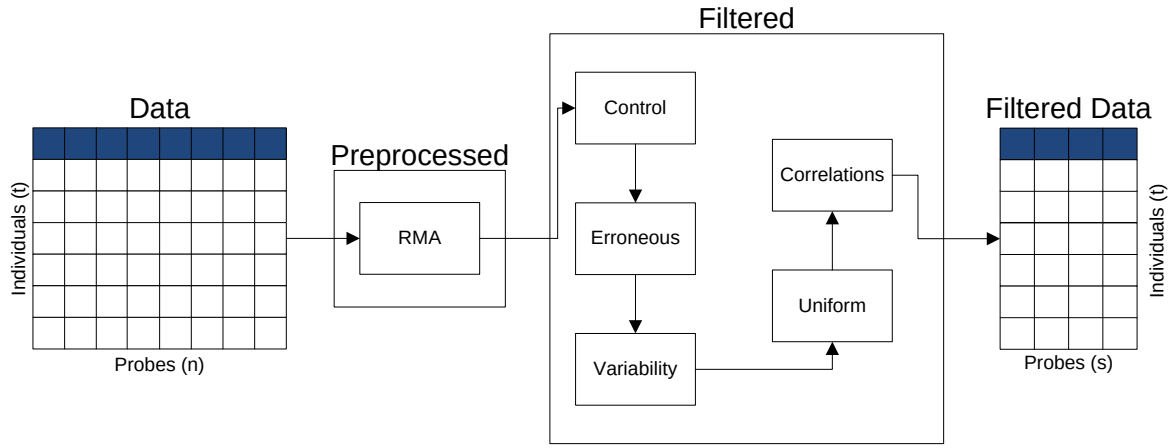
Table 2. Classification errors numerical

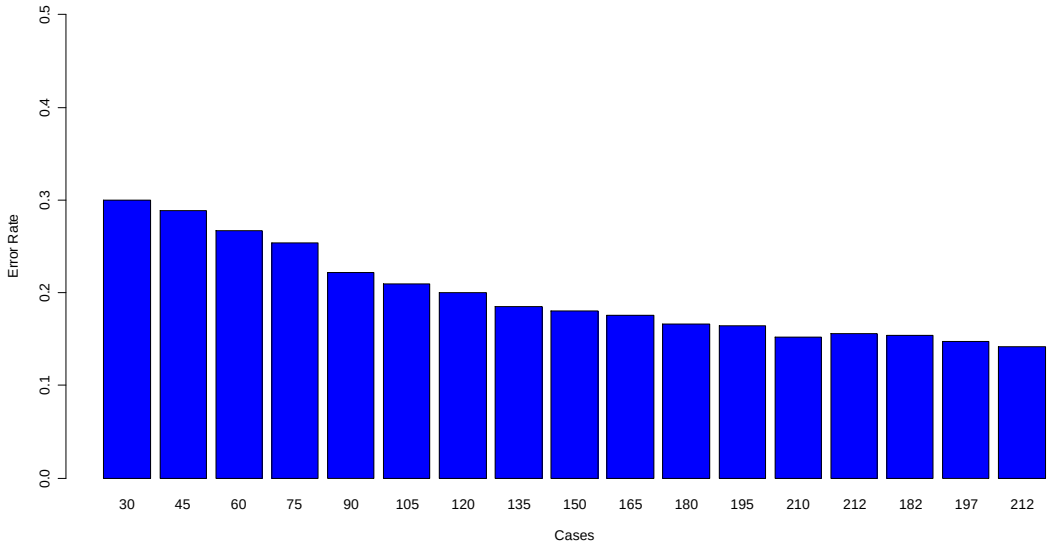
	Total	Error
ALL	10	3
AML	49	11
CLL	89	4
CML	22	7
MDS	42	5

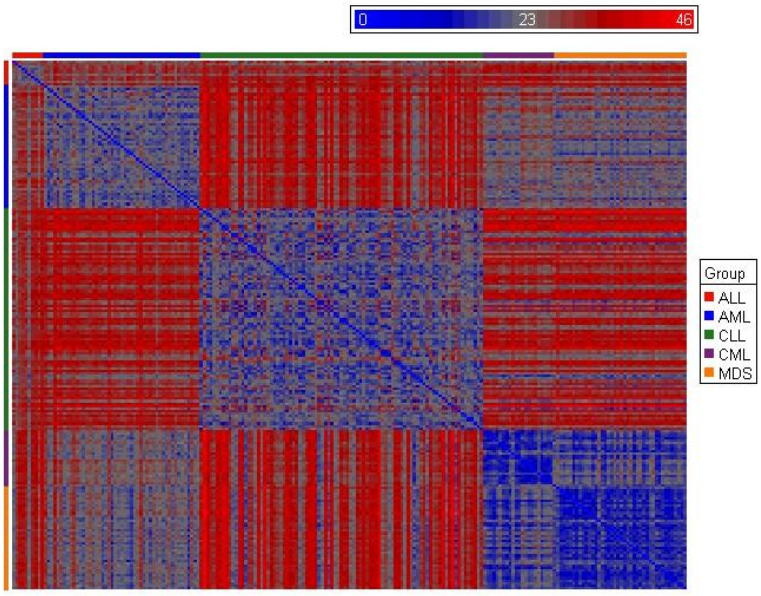
Table 3. Comparison of methods. * different median and = equal, (-) median of column less than median of row

	CBR	Dendrogram	PAM
CBR			
Dendrogram	*(-)		
PAM	*(-)	*(-)	

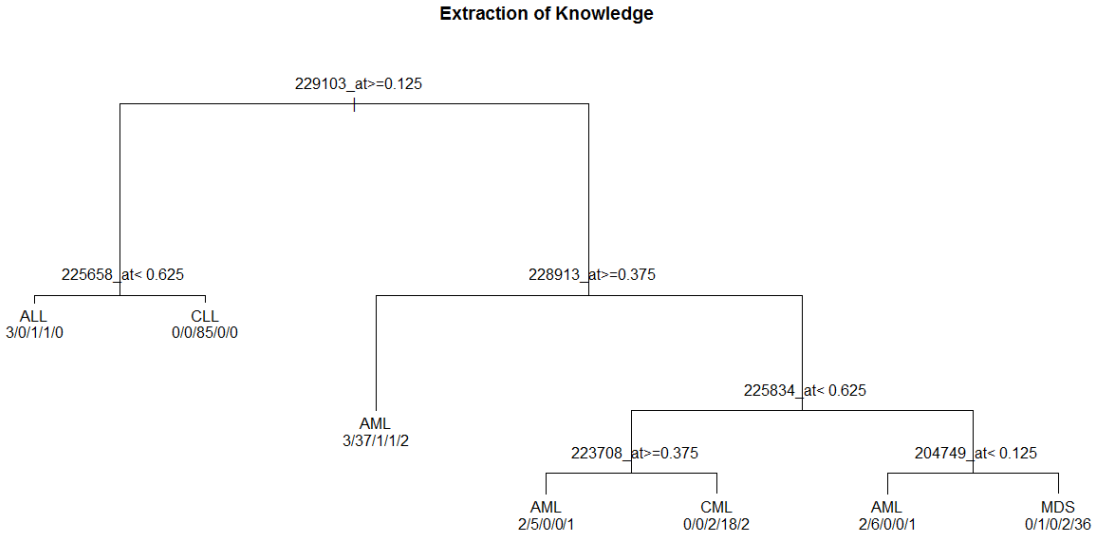

Retrieve
Laboratory
RMA
Control
Erroneous
Variability
Uniform
Correlation
$D^{'} = \{d_1^{'},...,d_i^{'}\}$
$d_i^{'} \in R^s, s \leq n$
Filtered Data
data
$D = \{d_1,...,d_i\}$
$d_i \in R^n$
Memory Cases
Reuse
Cluster Classification
Retain
new case
$d_{i+1}^{'} \in g_i$
$d_{i+1}^{'}$
Rules
$f_i = \{(z_1, v_1)...(z_i, v_i)\}$
$g_i = (f_1,..., f_i)$
$G = \{g_1,...,g_r\} r < s$
$g_i \subset D$
$g_i^{'} = g_i \cup \{d_{i+1}^{'}\}$
Cluster
CART
Expert
Revise

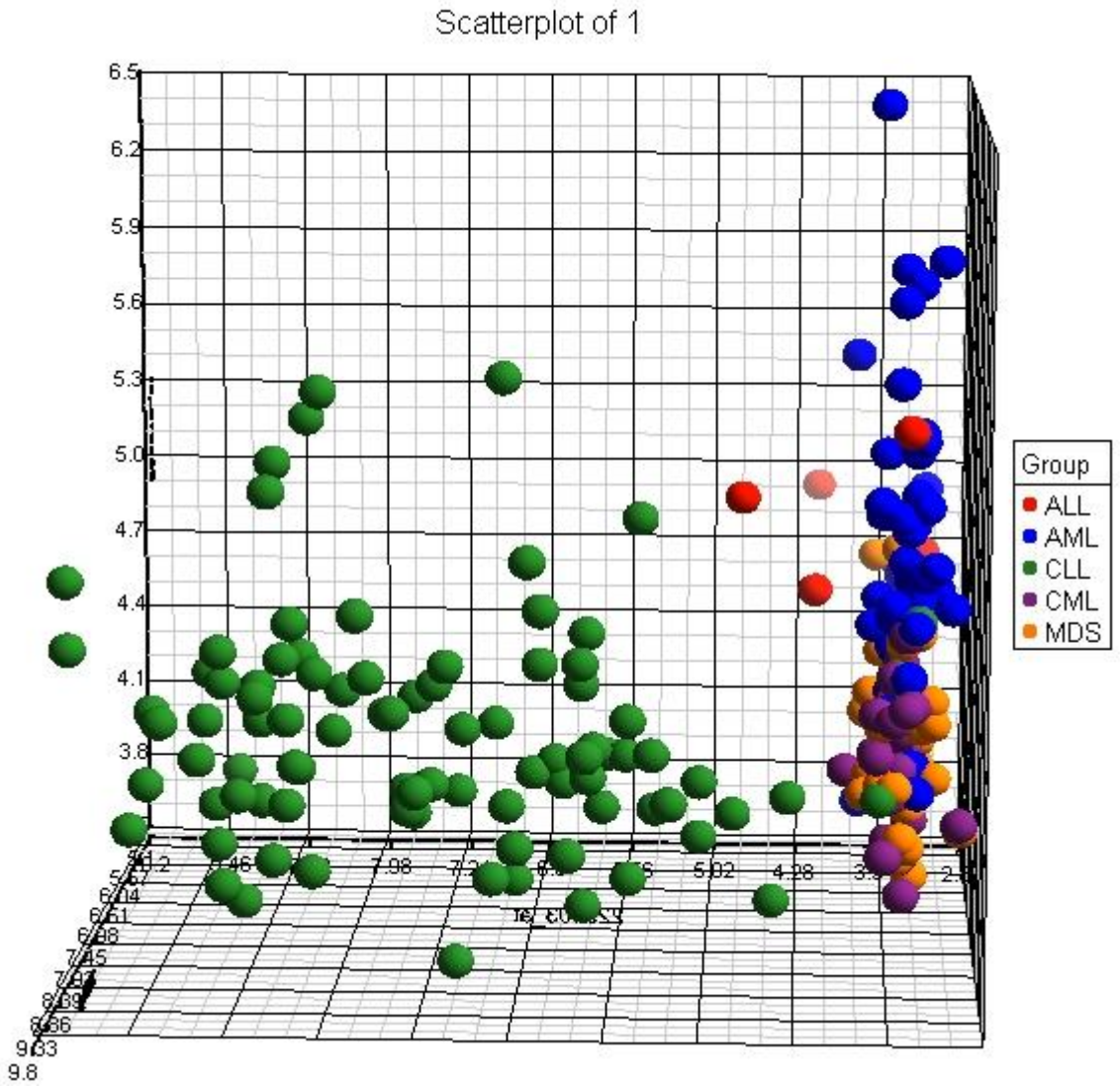

Data
Individuals (t)
Probes (n)
Preprocessed
RMA
Filtered
Control
Erroneous
Variability
Correlations
Uniform
Filtered Data
Individuals (t)
Probes (s)

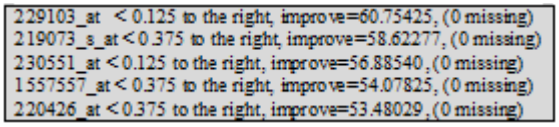

Error Rate
0.0
0.1
0.2
0.3
0.4
0.5
30
45
60
75
90
105
120
135
150
165
180
195
210
212
182
197
212
Cases

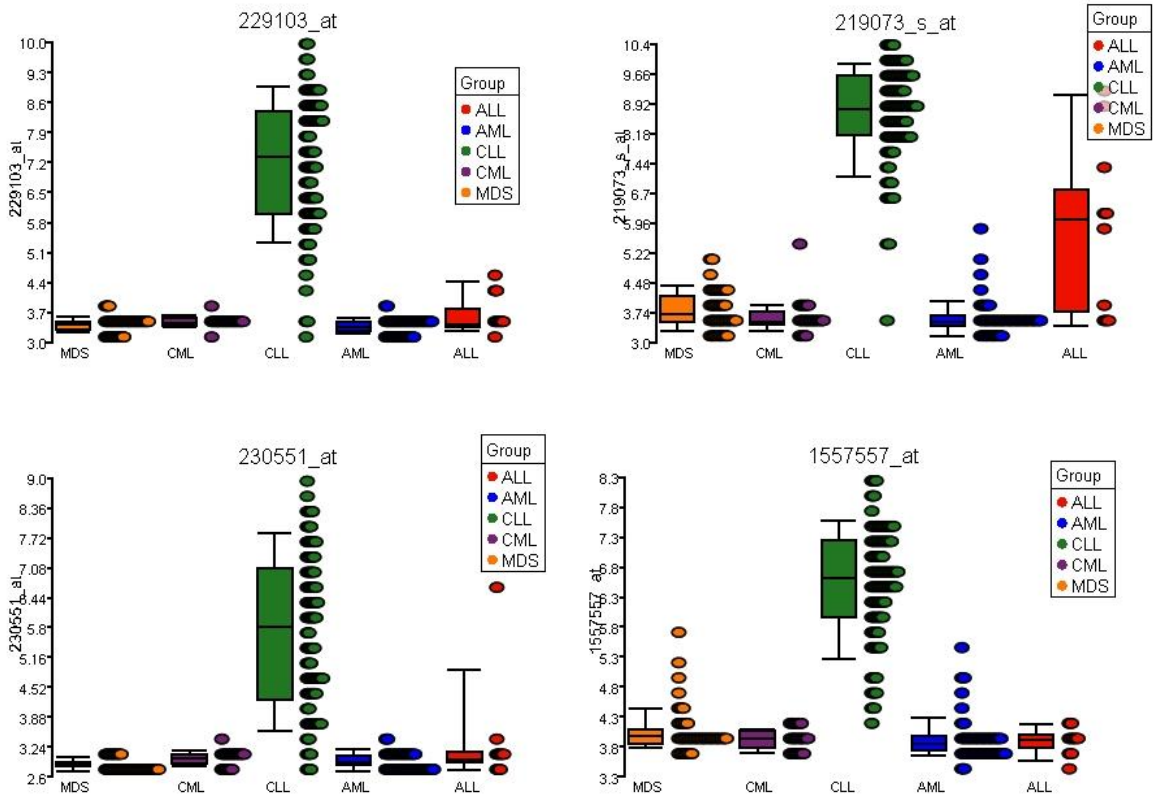

0
23
46
Group
ALL
AML
CLL
CML
MDS

```
1) root 212 123 CLL (0.047 0.23 0.42 0.1 0.2)
  2) 229103_at>=0.125 90   4 CLL (0.033 0 0.96 0.011 0)
    4) 225658_at < 0.375 4   1 ALL (0.6 0 0.2 0.2 0) *
    5) 225658_at >=0.375 86   0 CLL (0 0 1 0 0) *
  3) 229103_at< 0.125 122  73 AML (0.057 0.4 0.025 0.17 0.34)
    6) 228913_at228913_at>=0.375 44   7 AML (0.068 0.84 0.023 0.023 0.045)
     12) 218949_s_at>=0.625 4   1 ALL (0.75 0 0.25 0 0) *
     13) 218949_s_at< 0.625 40   3 AML (0 0.92 0 0.025 0.05)
       26) 1554783_s_at>=0.125 37   0 AML (0 1 0 0 0) *
       27) 1554783_s_at< 0.125 3   1 MDS (0 0 0 0.33 0.67) *
    7) 228913_at228913_at< 0.375 78  38 MDS (0.051 0.15 0.026 0.26 0.51)
     14) 225834_at< 0.625 30  12 CML (0.067 0.17 0.067 0.6 0.1)
       28) 223708_at>=0.375 8   3 AML (0.25 0.62 0 0 0.12)
         56) 1553787_at>=0.125 5   0 AML (0 1 0 0 0) *
         57) 1553787_at< 0.125 3   1 ALL (0.67 0 0 0 0.33) *
       29) 223708_at< 0.375 22   4 CML (0 0 0.091 0.82 0.091)
         58) 203074_at< 0.125 3   1 CLL (0 0 0.67 0 0.33) *
         59) 203074_at>=0.125 19   1 CML (0 0 0 0.95 0.053) *
     15) 225834_at>=0.625 48  11 MDS (0.042 0.15 0 0.042 0.77)
       30) 204749_at< 0.125 9   3 AML (0.22 0.67 0 0 0.11)
         60) 1552736_a_at< 0.125 6   0 AML (0 1 0 0 0) *
         61) 1552736_a_at>=0.125 3   1 ALL (0.67 0 0 0 0.33) *
       31) 204749_at>=0.125 39   3 MDS (0 0.026 0 0.051 0.92) *
```


Extraction of Knowledge
229103_at>=0.125
225658_at< 0.625
228913_at>=0.375
ALL
3/0/1/1/0
CLL
0/0/85/0/0
AML
3/37/1/1/2
225834_at< 0.625
223708_at>=0.375
204749_at< 0.125
AML
2/5/0/0/1
CML
0/0/2/18/2
AML
2/6/0/0/1
MDS
0/1/0/2/36


Scatterplot of 1
6.5
6.2
5.9
5.6
5.3
5.0
4.7
4.4
4.1
3.8
Group
ALL
AML
CLL
CML
MDS


229103_at < 0.125 to the right, improve=60.75425, (0 missing)
219073_s_at < 0.375 to the right, improve=58.62277, (0 missing)
230551_at < 0.125 to the right, improve=56.88540, (0 missing)
1557557_at < 0.375 to the right, improve=54.07825, (0 missing)
220426_at < 0.375 to the right, improve=53.48029, (0 missing)


229103_at
219073_s_at
230551_at
1557557_at
Group
ALL
AML
CLL
CML
MDS