Artificial Intelligence in Medicine (2009) **46**, 179—200

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Model of experts for decision support in the diagnosis of leukemia patients

Juan M. Corchado, Juan F. De Paz, Sara Rodríguez, Javier Bajo*

Departamento de Informática y Automática, Universidad de Salamanca, Plaza de la Merced s/n, 37008 Salamanca, Spain

Received 9 June 2008; received in revised form 11 November 2008; accepted 1 December 2008

KEYWORDS
Mixture of experts;
Enhanced self-organizing incremental neural network;
Leukemia classification

Summary

Objective: Recent advances in the field of biomedicine, specifically in the field of genomics, have led to an increase in the information available for conducting expression analysis. Expression analysis is a technique used in transcriptomics, a branch of genomics that deals with the study of messenger ribonucleic acid (mRNA) and the extraction of information contained in the genes. This increase in information is reflected in the exon arrays, which require the use of new techniques in order to extract the information. The purpose of this study is to provide a tool based on a mixture of experts model that allows the analysis of the information contained in the exon arrays, from which automatic classifications for decision support in diagnoses of leukemia patients can be made. The proposed model integrates several cooperative algorithms characterized for their efficiency for data processing, filtering, classification and knowledge extraction. The Cancer Institute of the University of Salamanca is making an effort to develop tools to automate the evaluation of data and to facilitate de analysis of information. This proposal is a step forward in this direction and the first step toward the development of a mixture of experts tool that integrates different cognitive and statistical approaches to deal with the analysis of exon arrays. The mixture of experts model presented within this work provides great capacities for learning and adaptation to the characteristics of the problem in consideration, using novel algorithms in each of the stages of the analysis process that can be easily configured and combined, and provides results that notably improve those provided by the existing methods for exon arrays analysis.

Material and methods: The material used consists of data from exon arrays provided by the Cancer Institute that contain samples from leukemia patients. The methodology used consists of a system based on a mixture of experts. Each one of the experts incorporates novel artificial intelligence techniques that improve the process of

* Corresponding author at: Escuela Universitaria de Informática, Universidad Pontificia de Salamanca, 37002 Salamanca, Spain. Tel.: +34 639771985/923 277100x7687; fax: +34 923 277101.

E-mail addresses: corchado@usal.es (J.M. Corchado), fcofds@usal.es (J.F. De Paz), srg@usal.es (S. Rodríguez), jbajope@usal.es (J. Bajo).

0933-3657/$ — see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.12.001

carrying out various tasks such as pre-processing, filtering, classification and extraction of knowledge. This article will detail the manner in which individual experts are combined so that together they generate a system capable of extracting knowledge, thus permitting patients to be classified in an automatic and efficient manner that is also comprehensible for medical personnel.
Results and conclusion: The system has been tested in a real setting and has been used for classifying patients who suffer from different forms of leukemia at various stages. Personnel from the Cancer Institute supervised and participated throughout the testing period. Preliminary results are promising, notably improving the results obtained with previously used tools. The medical staff from the Cancer Institute considers the tools that have been developed to be positive and very useful in a supporting capacity for carrying out their daily tasks. Additionally the mixture of experts supplies a tool for the extraction of necessary information in order to explain the associations that have been made in simple terms. That is, it permits the extraction of knowledge for each classification made and generalized in order to be used in subsequent classifications. This allows for a large amount of learning and adaptation within the proposed system.

1. Introduction

During the last few years, there have been great advances in the field of Biomedicine [1]. The incorporation of computational techniques and artificial intelligence in medicine has led to notable progress in the prevention and detection of diseases [1]. One of the areas of medicine that is in its height of development and is fundamental in the application of techniques that facilitate the automatic treatment of data and the extraction of knowledge is genomics. Genomics involves the study of genes, their genetic sequencing, structure and relationship [2]. There are four distinct fields within the study of genomics. One of them is transcriptomics, which deals with the study of messenger ribonucleic acid (mRNA) using techniques such as expression analysis [3]. These techniques study the ribonucleic acid (RNA) strands by identifying the expression level for each of the genes studied. They consist of the exposure of the DNA molecules to complementary DNA (cDNA) molecules obtained from messenger RNA (mRNA). The mRNA molecules are marked with different bold dyes. The DNA and cDNA molecules are matched by pairs. In this process, the cDNA molecules that are not paired with any gene will be eliminated from the microarray. Finally, using a scanner, an image of the microarray is obtained by measuring levels of color. The different levels of fluorescence obtained can be analyzed and represented as a data array. The methods and tools traditionally used were developed to work with expression arrays that contain approximately 50,000 data points. However, the emergence of exon arrays [4] denotes an important breakthrough in biomedicine. The exon arrays need new tools and methods that can work with quantities of up to 5,500,000 data points.

This study presents a system that is consistent with a mixture of experts model that facilitates the analysis and classification of data obtained from exon arrays from leukemia patients. Leukemia, or blood cancer, is a disease that has a high cure rate with early detection [5]. The proposed system within the context of this study focuses on the detection of cancerous patterns found in the data extracted from the exon arrays taken from patient samples provided by the Cancer Institute of the University of Salamanca. The system provides suggestions about the classification of leukemia patients and represents the analysis process by means of rules. Through these rules the medical staff can extract knowledge about the entire classification process, including the decisions taken by each of the expert models. It is assembled from a selection and mixture of expert systems considered optimal for use in each of the following stages: (i) pre-processing and filtering of data, (ii) the application of clustering techniques, and (iii) the extraction of knowledge from an expression analysis. The selection of each one of these expert systems was made by considering the characteristics of the data corresponding to the leukemia patients, and was validated by comparing against other techniques and methods used for resolving problems with similar characteristics. The proposed model integrates several cooperative algorithms characterized for their efficiency for data processing, filtering, classification and knowledge extraction. The Cancer Institute of the University of Salamanca is making an effort to develop tools to automate the evaluation of data and to facilitate de analysis of information. This proposal is a step forward in this direction and the first step toward the development of a mixture of experts system that integrates different cognitive and statistical approaches to deal with the analysis of exon arrays.

Exon arrays are chips with a significantly higher number of functions compared to their predecessors [6]. The characteristics of exon arrays allow for a large number of data to be analyzed and classified for each patient (approximately 5.5 million features per array). However, the high dimensionality of data produced by an exon array makes it impossible to use the majority of the previously employed techniques for expression array analysis (which contain approximately 50,000 probes), and calls for the development of new techniques and tools. The high dimensionality of data supplied by each exon array presents problems in handling and processing, thus making it necessary to improve each of the steps of expression array analysis in order to obtain an efficient method of classification. An expression analysis basically consists of three steps: normalization and filtering, clustering and classification, and extraction of knowledge. These steps can be automated and included within an expert system. The first step is fundamental for achieving a good standardization of data, and a preliminary filtering process to reduce the dimensionality of the set of data used [7]. Since the problem at hand deals with high dimensional arrays, it is important to have a very good pre-processing technique that can facilitate automatic decision-making with regards to selecting the most vitally important variables for the classification process. In light of these decisions, it will be possible to reduce the set of original data. Additionally, the selection of a clustering technique in the second phase of analysis allows the data to be grouped according to certain variables that control the behavior of the group [8]. After the organization of groups, patients can be classified and assigned into the group with which they share the most similarities. Finally, an extraction of knowledge system facilitates the interpretation of the results obtained after the pre-processing and classification steps, thus making it possible to learn from the information acquired from the results [9]. The process of extracting knowledge shapes the knowledge obtained into a set of rules that can be used for improving new classifications [9].

The system proposed in this study presents a novel synthesis that encompasses various fields of artificial intelligence (filtering techniques, clustering, artificial neural networks, and extraction of knowledge). Specifically, the system presented in this article uses a model that combines the advantages of three novel methods for the analysis of data from exon arrays. The default Affymetrix background correction and robust multi-array average (RMA) are improved by means of novel algorithms that are used for a better data pre-processing, filtering and reduction in the dimensionality of the data. These techniques also eliminate those data that do not contribute to the classification process. An enhanced self-organizing incremental neuronal network (ESOINN) clustering technique [10] has also been integrated in the proposed model. It allows both the incorporation of the distribution process along the entire surface of classification, and the separation into low density groups. Finally the classification and regression trees (CART) has been included in the tool and identified as an excellent knowledge extraction technique [11] that facilitates the observation and study of the completed classification process, as well as the deduction of rules that can be applied for improving subsequent analyses. The proposed model introduces significant improvements in the analysis process, leading to an increased success rate in the classification of patients and a decrease in the number of false positives. The proposed model facilitates the analysis of data in an automatic way, providing learning and adaptation capacities, and integrates several cognitive techniques in a way that has never been done. The main advantage of the development of the tool is that facilitate the treatment of a huge amount of data in a simple and supervised way and that provides together with the solution an explanation in the form of rules. In this way artificial intelligence is not any more a black box and the users of the tool will be able to follow how the solution to a problem is constructed.

The article is organized as follows: first we present a description of the problem that that instigated this research: the classification of patients suffering from cancer of the blood based on samples obtained from exon arrays. Traditional techniques used in each phase of data analysis such as data preprocessing, data clustering, and extraction of knowledge are outlined in this section. Section 3 presents the proposed model consisting of a mixture of experts for the analysis and classification of data corresponding to leukemia patients. Section 4 shows the results obtained from the proposed model and compares them with the results obtained by using other techniques. Finally, Section 5 describes the conclusions obtained from the given results.

2. Computational methods in the investigation of cancer

Hematological cancers such as leukemia have been the object of genetic and chromosomal analyses for many years [12—14]. The relationship between chromosomal alterations and the prognosis of leukemia and lymphomas are well situated within this field of study. Recently, conventional studies on expression arrays have demonstrated that chromosomal alterations are associated with characteristic patterns of expression. Leukemia is a type of blood cancer that results from an abnormal functioning of the bone marrow, which tends to cause an abnormal proliferation of white or red blood cells [15]. The four most important types of leukemia are: acute and chronic myeloid leukemia (AML, CML), and acute and chronic lymphocytic leukemia (ALL, CLL) [15]. Although there have been extensive studies on the subject, the actual cause of leukemia continues to be a mystery. Nearly 25,000 new cases of acute and chronic leukemia appear every year. The majority of those cases occur in adults and people approximately age 60 and older, but the number of cases of acute lymphocytic leukemia in children has increased over the last few years. Each year approximately 10,000 adult cases are diagnosed as AML, 8000 as CLL, 500 as CML, and 3500 as ALL [16]. The rest are unclassified types of blood cancer. A recent study [17], surveillance epidemiology and end results, 2007, shows an estimated 19,900 new cases of myeloma diagnosed in the United States in 2007.

Microarrays have been successfully tested in identifying leukemia prognoses. They have become an essential tool in genomics research, making it possible to investigate the global genetic expression in all aspects of human disease [1,18]. In the exon arrays, the information is divided into probe selection regions (PSRs). The PSRs are contiguous and do not overlap in genomic space, they are grouped in exon clusters and in turn, these are grouped into transcript clusters. Finally, a transcript cluster roughly corresponds to a gene. Subsequently, the data from the PSR are used for measuring the expression of a particular gene by means of fluorescent intensities. This process of studying microarrays is called expression analysis and consists of four phases: obtaining data, pre-processing data, statistical analysis and biological interpretation.

The development of microarray technology can generate enormous amounts of data. For this reason, data mining and machine learning techniques have been aptly used in incipient areas of investigation that have resulted from microarray analysis [19]. Additionally, a new generation of microarrays was recently designed with a much greater wave density than was previously available, thus allowing the waves to be organized in exons. As a result, the study of gene expression can be performed in much greater detail than ever before. Affymetrix GeneChip microarray, one of the most popular organizations for measuring gene expression, has released this new generation of microarrays, exon arrays, designed to interrogate exon-level expression [6]. Exon arrays, as can be seen in Fig. 1 (modified from Affymetrix exon array design datasheet [20]), differ significantly from their predecessor arrays (3′ expression arrays) in the number and placement of the oligonucleotide probes and in the design of control probes for background correction. The Affymetrix human exon 1.0 ST array contains approximately 5.5 million probes, forming 1.4 million probe sets [4]. Thus, there are about six times as many features as in the previous generation of chips.

Exon arrays contribute to the dramatic increase in data available to improve the quantitative estimation of gene-level expression. However, it is necessary to develop new techniques capable of working with data provided by the exon arrays. Previous research related to the expression analysis of exon arrays is outlined below. The main techniques used in the analysis of exon arrays are briefly explained in the following paragraphs, focusing on the lacks of the existing techniques and the advantages provided by our proposal.

Prior to analyzing microarray 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 currently exists a limited group of algorithms that investigators use for performing these steps. The most common are Affymetrix

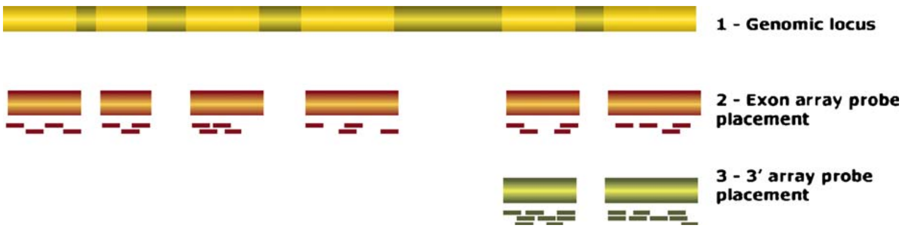


Figure 1 Exon array structure. Probe design of exon arrays. (1) Exon—intron structure of a gene. Gray boxes represent introns, rest represent exons. Introns are not drawn to scale. (2) Probe design of exon arrays. Four probes target each putative exon. (3) Probe design of 3′expression arrays. Probe target the 3′end of mRNA sequence.

microarray suite 5.0 (MAS5.0) [20], probe logarithmic intensity error (PLIER) [21], and RMA [22]. Table 1 shows each of their most important characteristics. As can be seen in Table 1, MAS5.0 is suitable for single arrays and is less sensitive than PLIER and RMA to small changes in the data. RMA is frequently used because it minimizes the variance seen across the arrays, however, PLIER is preferred when it is important to detect fold changes. Once the data have been pre-processed, it is necessary to filter the information obtained so that it will be possible to reduce the number of probes and facilitate the execution of the subsequent experts. The technique chosen for the model proposed in the framework of this investigation is a combination of statistical tests.

After the pre-processing and filtering the next step in the analysis process is to perform the clustering of individuals based on their proximity according to their probes. Since the problem on which this study is based contained no prior classification with which training could take place, a technique of unsupervised classification was used. There is a wide range of possibilities. Some of these techniques are artificial neural networks such as self-organizing map (SOM) [23], growing neural gas [24] resulting from the union of techniques, competitive Hebbian learning (CHL) [25] and neural gas (NG) [26], growing cell structure (GCS) 0, growing grid or the self-organizing incremental neuronal network (SOINN) [27]. Some of the methods, such as self-organized Kohonen maps, set the number of clusters in the initial phase of training when using the algorithm of the k-means learning method. This is the reason that 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 varied and the degree of waste compaction checked so that according to this value, the final number of groups could be set. 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 behavior 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 extend, the adaptation of the mesh. Unlike the SOM 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. After analyzing different techniques and checking the problems they might present so that they might be applied to the problem at hand, we have decided to use a variation of SOINN [27], called ESOINN [10] in the proposed model.

Finally, once the clustering has finished, it is important to learn from the classification obtained. 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 are provided to support the decision-making process. It should be noted that extraction of knowledge 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.

Among all of the techniques used for extraction of knowledge, the most useful are those within the field of machine learning. A familiar component in this area is the neural networks, from which a number of applications have been developed with satisfactory results. The use of neural networks for discovering clusters within the data presents the occasional problem of trying to extrapolate a meaning for every grouping. There are two ways of solving this problem. The first refers to obtaining the median value for each characteristic within the data. The second requires passing each cluster through a machine learning algorithm in order to generate a set of rules that describe the characteristics of that grouping [28]. An alternative to the neural networks can be found by focusing on rough sets [29]. The rough sets theory assumes that knowledge can be represented in a decision table. The output produced by this method depends in large part on the attributes or variables of elements of the universe, and the values that the attributes can have. If the

Table 1 Comparison of traditional pre-processing techniques.

	Advantage	Disadvantage
MAS5.0	Single-array algorithm is independent of other data in data set	Not as sensitive as either RMA or PLIER to small changes in target abundances
PLIER	Ability to detect small fold changes	More variance in individual signals than seen with RMA
RMA	Minimizes the variance seen across the arrays	Compresses fold changes for low-intensity probe sets

number of attributes is too high, as is the case study presented in this work, the output decreases considerably. Furthermore, by analyzing the functioning of this theory, it is possible to see that it is geared more towards qualitative than quantitative variables. Nevertheless, the methods that offer the best results for extraction of knowledge are the rule induction and decision tree systems. Examples of those that have made a significant impact within the field of biomedicine include concept learning system [30], induction decision trees [31], CART [11], oblique classier 1 [32], ASSISTANT [33], and C4.5, C5.0/See5 [34]. For example, it is possible to find applications of the previously mentioned algorithms in extraction of knowledge that can help to make predictions from the analysis of genes [35].

In this study we have concentrated on the methods that use induction rule and decision tree algorithms since they more easily adaptable to the characteristics of the problems we are attempting to solve. The CART algorithm was chosen as the technique to apply to the model proposed in this study because of its wide acceptance and proven efficiency in extraction of knowledge. This type of classification system presents important advantages [11,36] for its application within the realm of bioinformatics. Some of them include:

- It does not depend on the distribution of dependent and independent variables since it is a non-parametric method.
- Variables do not need to be independent.
- The variables can be qualitative, quantitative or a combination of both.
- It lends itself to working with a large number of variables in an efficient matter.
- It is superior to expert systems in that human expert intervention is not necessary for the inference of classification rules, since these are automatically generated.
- It has the advantage over neural networks in that the rules that are generated are much more comprehensible for the user than the network interface topology. A neural network, for however easy it may be, follows the black box model which does not value the relative importance of each of the explicative variables.
- It allows the use of probability values for classifying individuals, as shown in the results presented in Section 5 of this article.
- It is ''independent'' from the transformation of independent variables.
- It handles atypical values in an efficient manner.

The next section introduces the model proposed in this study, which incorporates the mixture of experts model presented in the current section. Additionally, we present a method in which the mixture of experts can be integrated to obtain a model that allows leukemia patients to be classified in an efficient manner.

3. Mixture of experts model

The proposed model, that incorporates the mixture of three experts in sequential form, is presented in detail in this section. This model has the advantage of integrating different techniques, in a novel way, considered to be optimal for using in the stages of the expression analysis for the problem of classifying leukemia patients. This way, the techniques that offer good results in each phase are combined to obtain the most optimal result overall. The proposed model considers the characteristics of each expert in order to achieve an appropriate integration.

The proposed model for the system consists of generating a series of independent procedures that are based on different paradigms that are performed by a series of independent subtasks. These procedures are distributed over various modules thus making it possible to integrate the different experts. Because the experts can now communicate directly, it is possible to generate a global result. The proposed model within the scope of this investigation can be divided into three modules that will each be responsible for carrying out a set of tasks, specifically: pre-processing/filtering, clustering, and extraction of knowledge. Additionally, another module is incorporated to represent the information and present the results to medical personnel in a comprehensible manner. The structure of these modules can be observed in Fig. 2. Fig. 2 represents the modules as arrows with their corresponding assigned task: pre-processed/filtered, clustering, extraction of knowledge and representation. Each of the modules receives certain input data and provides output data in a fixed format. As can be seen in Fig. 2, the different modules work independently in order to facilitate the modification of any of the proposed experts, or to incorporate simple new techniques (including new experts). This incorporation would only affect the expert of a single module, while the others remain unchanged. This allows a generalization of the model so that problems with different characteristics can be studied, thus facilitating the introduction of different experts in each module and making it possible to select the expert best suited to apply in each particular problem.

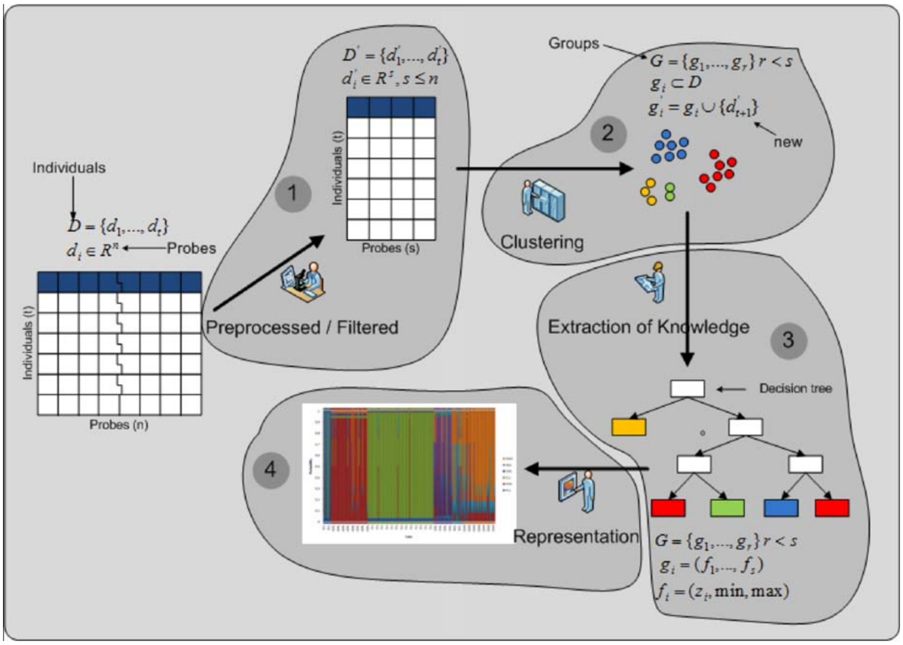


Figure 2 Proposed expert model. The initial problem description is composed of all the individuals (t) together with the n probes. The first expert pre-processes and filters the probes, reducing the set of probes to s elements but maintaining the t individuals. The second expert executes the clustering, creates r groups and assigns the new individual (t + 1) to one of these groups. The third expert explains how the individual elements have been classified into groups by means of a knowledge extraction technique, and by obtaining a graphical representation (a tree). The final module represents the probability of assigning individuals to each of the groups depending on the probes selected, taking into account the knowledge extracted.

Fig. 2 shows a scheme of the bio-inspired model intended to resolve the problem described in Section 2. The proposed model follows the procedures that are performed in medical centres. Data is first pre-processed and filtered using the improved Affymetrix background corrections and RMA methods, the data clusters are identified using an ESOINN network and finally the knowledge extraction is carried out with a non-parametric statistical method called CART. As can be seen in Fig. 2, a previous phase, external to the model and critical in the expression analysis, consists of a set of tests which allow us to obtain data from the chips and are carried out by the laboratory personnel. The chips are hybridized and explored by means of a scanner, obtaining information on the marking of several genes based on the fluorescence. At that point, the model starts to process the data obtained from the exon arrays. The system receives a huge amount of data D, which needs to be reduced in order to apply classification techniques. In the pre-processing/filtering module the data is pre-processed to eliminate defective samples and get standardized measures. Moreover, the pre-processed data is filtered to reduce the dimensionality and to eliminate those that do not contribute, or do so insignificantly, any value to the classification process. The result of this module is represented as D', as shown in Fig. 2. D' is then used as the input for the clustering module, which uses the ESOINN algorithm to classify the data contained in D' into groups represented as G. Then, the new individual is presented to the network which assigns it to the group g_i. The groups obtained and the new classification are then studied by the extraction of knowledge module to update the set of rules used to support the decision making process. Finally, Fig. 2 shows how the results provided by the extraction of knowledge module are sent to the representation module to present the data in a comprehensible format to the user. The next sub-sections will describe in detail the structure of the three experts used to construct de analysis model.

3.1. Pre-processing and filtering: improved Affymetrix background correction and RMA

The default Affymetrix background correction and RMA techniques has been used for normalization and summarization. We have chosen RMA method as a "best practice" owing to its wide use and because we are highly confident of RMA based on previous experiences. We want to stress that the focus of our experiment is primarily to prove the feasibility of new clustering methods rather than optimizing the pre-processing phase. RMA 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, the goal of which is to make the distribution of probe intensities the same for arrays; and (iii) expression calculation: performed separately for each probe set n. To obtain an expression measure we assume that for each probe set n, the background adjusted, normalized and log transformed intensities, denoted with Y, follow a linear additive model

$$x_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn} \text{ with } i = 1, \ldots, I,\ j = 1, \ldots, J,\ n = 1, \ldots, N,\ \sum_{j} \alpha_{j} = 0 \tag{1}$$

where α_j is a probe affinity effect, μ_i represents the $\log_2$ scale expression level for array i and ε_{ij} represents an independent identically distributed error term with mean 0. Median polish 0 is used to obtain estimates of the values.

Once the pre-processing phase has been completed, the filtering phase begins. The purpose of this initial phase is to significantly reduce the dimensionality of the data and to eliminate those that do not contribute, or do so insignificantly, any value to the classification process. This can be considered the recovery phase of important variables needed for the classification process. This study proposes a novel filtering method that is separated into the 5 stages that will be described in the following sections.

3.1.1. Control

During this phase, all probes used for testing hybridization are eliminated. These probes have no relevance at the time when 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.

3.1.2. Errors

On occasion, some of the measures 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.

3.1.3. 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

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

where N is the number of items total, $\bar{\mu}_{\cdot j}$ is the average population for the variable j, x_{ij} is the value of the probe j for the individual i.

2. Standardize the above values

$$z_{i} = \frac{\sigma_{\cdot j} - \mu}{\sigma} \tag{3}$$

3. Discard of probes for which the value of z meet the following condition: $z < -1.0$ given that $P(z < -1.0) = 0.1587$. This will effect the removal of about 16% of the probes if the variable follows a uniform distribution.

3.1.4. 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 followed is explained below, using the Kolmogorov—Smirnov [38] test as an example. H_0: The data follow a uniform distribution; H_1: The analyzed data do not follow a uniform distribution. Statistical contrast:

$$D = \max\left\{D^{+}, D^{-}\right\} \tag{4}$$

where

$$D^{+} = \max_{1 \le i \le n}\left\{\frac{i}{n} - F_{0}(x_{i})\right\}, \quad D^{-} = \max_{1 \le i \le n}\left\{F_{0}(x_{i}) - \frac{i-1}{n}\right\}$$

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:

$$D_{\alpha} = \frac{C_{\alpha}}{k(n)} \tag{5}$$

in the special case of uniform distribution $k(n) = \sqrt{n} + 0.12 + (0.11/\sqrt{n})$ and a level of significance $\alpha = 0.05$ $C_\alpha = 1.358$ $\alpha = 0.1$ $C_\alpha = 1.224$.

3.1.5. 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.

$$r_{x_{.i}y_{.j}} > \alpha \tag{6}$$

being $\alpha = 0.95$, where $\sigma_{x_{.i}x_{.j}}$ is the covariance between probes *i* and *j*.

Once filtered and standardized, the probes produce a set of values x_{ij} with $i = 1,\ldots, N$, $j = 1,\ldots, s$ where *N* is the total number of cases, *s* the number of end probes. The proposed filtering process is very novel since we are not aware of any previous research that uses a filter with the same characteristics as those found in this type of problem.

The pre-processing and filtering technique helps to notably reduce the dimension of the initial dataset, which is one of the main problems when working with exon arrays. At this moment the data is ready to be classified using clustering techniques. The expert system proposed in the framework of this research proposes an ESOINN neural network as a novel clustering method for exon arrays. The ESOINN clustering technique is explained in detail in the next section.

3.2. Clustering technique: ESOINN

ESOINN is the network selected as the second expert. It 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. This process is conducted in an iterative way until it reaches stability. Only the changes in the training phase are detailed below:

(a) 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 [39].

$$\begin{aligned} \Delta W_{a_1} &= n_1(M_{a_1})(\xi - W_{a_1}) \\ \Delta W_i &= n_1(M_{a_i})(\xi - W_{a_i}) \text{ with } i \in N_i \end{aligned} \tag{7}$$

The new learning tasks included in the algorithm are $n_1(x) = 1/\sqrt{x}$, $n_2(x) = 1/\sqrt{2 + x^2}$.

(b) 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

$$z_i = \frac{e_i - \mu}{\sigma}, z \equiv N(0,1) \text{ then } f(z) = \frac{1}{\sqrt{2\pi}} \text{Exp}\left[\frac{-z^2}{2}\right] \tag{8}$$

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.

The new algorithm defines an automatic threshold to automatically remove the connections.

(c) If all input patterns have been passed then a KS-Test [38] 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. The algorithm incorporates a new method is defined to automatically decide when the classification process should be finished.

Once, the clusters have been made, the new sample is classified. Assignments are made based on the k-nearest neighbour (KNN) algorithm [40] that allows values of probability to be established for each neighbour, for which it is necessary to establish a measure of similarity to calculate the distance between individuals. The similarity measure used is as follows:

$$d(n,m) = \sum_{i=1}^{s} f(x_{ni}, x_{mi}) w_i \tag{9}$$

where *s* is the total number variables, *n* and m the cases, w_i the value obtained in the uniform test and *f* the Minkowski distance [41] that is given for the following equation:

$$f(x,y) = \sqrt[p]{\sum_i |x_i - y_i|^p} \text{ con } x_i, y_i \in R^p \tag{10}$$

This dissimilarity measure weighs those probes that have a less uniform distribution, since these variables do not allow a separation. In order to validate the selected distance, the Kruskal—Wallis [42] test was carried out. It was verified whether the proportion of errors in the classification of each one of the individuals were the same for each group, bearing in mind the different measure. The results are shown in Table 2. Table 2 shows the non-para-

Table 2 Comparison of functions of distance. The table shows a comparison of equality of medians for the different functions of distance. The variable considered is the number of classification errors in each of the groups. The comparison is based on the Kruskal—Wallis non-parametric test.

	Minkowski	Euclidean	Max absolute
Minkowski			
Euclidean	=		
Max absolute	*(−)	*(−)	

metric Kruskal—Wallis test for independent group comparison taking into account their proportion of errors. The test compares if the samples come from the same population, then the variable is the proportion of errors for each group obtained in the classification, taking into account different distance metrics, is the same for the different methods. In this way, it is possible to determine if both methods can be considered as similar for a given confidence level (in this case the confidence level is stated at 0.05). If the result obtained shows different values for the proportion, this fact is represented as (−) in the table, and if the proportion of errors in a column is lower than the proportion of errors for the element in the row, and (+) otherwise. As can be seen in Table 2, the worst results are obtained for the absolute distance. In the hypothesis contrast the null hypothesis has been rejected (samples come from the same population) *, so the groups have different averages. Moreover, the proportion of errors is minor (−). On the other hand, the Minkowski and Euclidean distances provides similar results, and is represented as =.

The clustering technique allows us to classify leukemia patients and assign them to the groups represented as clusters. However, it is of interest to analyze the classification process followed to extract conclusions that can help to make more accurate diagnosis in the future. In this sense, it is necessary to use extraction of knowledge techniques, as explained in the following section.

3.3. Extraction of knowledge techniques: CART

Knowledge extraction is especially important when complex algorithms that use hard computing techniques and that generate models in an automatic way are used. Human experts are much confident when they know exactly why or at least how a solution to a problem has been calculated. CART is a non-parametric statistical method for extraction of knowledge in classifications. The extracted information is represented in a binary decision tree, which allows individuals to be classified from the root node. Keeping the kind of dependent variable in mind, CART can be separated into two types: classification tree, if the dependent variable is categorical; and regression tree in the case of a continuous dependent variable. For example, in the case study presented in Section 5, only the classification tree is useful since the dependent variable will be the patient's type of illness. The algorithm used for creating the decision tree used by CART is shown below:

(a) Define the impurity function $i(t)$ for each variable/wave where t is the current node. In this case, the Gini impurity function [11,36] was selected since it is more widely used. There are others such as Twoing [36] which produce more balanced trees, but the output follows an exponential pattern as the number of groups increases. Furthermore, by observing the results, it is easy to see that the tree created by the Gini index is not complex enough to propose the Twoing alternative.

$$i_v(t) = 1 - \sum_{j=1}^{r} p^2(j|t) \tag{11}$$

where $p(j|t)$ is the frequency relative to class j in node T, and vs is the selected variable $v = 1, \ldots, r$.

(b) Calculate the value of the split s in each variable v from node t that maximizes the expression (12).

$$\Delta i_v(s,t) = i(t) - p_l \cdot i(t_l) - p_r \cdot i(t_r) \tag{12}$$

where p_l is the number of cases that end up at the left son node for t p_r those that end up on the right, t_l is the left node and t_r the right node.

(c) Select the greater of $\Delta i_v(s,t)$ among all variables v.

(d) Repeat from step 1 for all nodes with no children having more than one class of elements $\Delta i_v(s,t) = 0$ for all variables v.

Once the decision tree for classifying individuals according to wave values has been formed, the CART algorithm is applied for the pruning phase, which consists of eliminating nodes to reduce the complexity and improve the tree's ability to generalize [36]. A minimal cost complexity pruning function is established for the pruning phase. The cost function depends on the complexity of the tree (number of leaf nodes n). The error rate for the tree T $R(T)$ and α is the parameter for complexity.

$$R_\alpha = R(T) + \alpha \cdot n \tag{13}$$

At this point the data is modified in order to optimize the functioning of the CART algorithm. By analyzing the behavior of the algorithm, it is easy to note that the number of different levels for the continuous variables affects their output. Because of this, the discretization process is applied to the values corresponding to the continuous variables that are available. The discretization of the values allows the efficient generation of the decision tree. Our proposal incorporates such discretization to reduce the processing costs. Otherwise, it was impossible to work with such a large number of continuous variables. There are 5 levels selected during discretization, which allows the values for fluorescence to be represented as: very low, low, medium, high, and very high. The following section explains the process that was followed, including fuzzy logic criteria:

(a) Select the maximum value M_j and the minimum value m_j for each variable with $j = 1,\ldots, r$ where r is the number of variables.

(b) Transform the data for x_{ij} as follows:

$$x_{ij} = \frac{x_{ij} - m_j}{M_j - m_j} \tag{14}$$

(c) Assign a fuzzy value to the data according to the following equation

$$x_{ij} = \alpha \quad \text{if}\, \alpha - \beta/2 < x_{ij} \leq \alpha + \beta/2 \tag{15}$$

where $\alpha = k/n$ with $k = 1,\ldots, n$, n is the number of intervals and β the amplitude of intervals.

Finally, in order to avoid overloading the algorithm, we eliminate all variables with a median value equal for all groups by using the Kruskal—Wallis [42] non-parametric test for equality of medians. These variables will not allow a classification of individuals since the values for the individuals from different groups are intermixed. We can then conclude that the median can be considered equal for all individuals. Once the discretization for variables has been applied and all the variables with a null hypothesis from the previous contrast have been eliminated, the procedure is completed with the previous process of adaptation of data preceding the application of extraction of knowledge. At this time, the extraction of knowledge process can be applied by using the previously mentioned CART method.

3.4. Working model

The expert responsible for carrying out the preprocessing/filtering of data first receives information from the laboratory on the various hybrid genes and the fluorescence values from the exon arrays that were assigned to each of the individuals or patients that were subjects in the test. This phase receives an array with a patient's data as input information. It should be noted that there is no filtering of the patients, since it is the work of the researcher conducting this task. The step filters genes but never patients. The aim of this phase is to reduce the search space to find data from the previous cases which are similar to the current problem. The set of patients is represented as $D = \{d_1,\ldots, d_t\}$, where $d_i \in R^n$ represents the patient i and n represents the number of probes taken into consideration. As explained in Section 3.1, during the pre-processing phase the data are normalized by the RMA algorithm [22] and the dimensionality is reduced bearing in mind, above all, the variability, distribution and correlation of probes. The result of this phase reduces any information not considered meaningful to perform the classification. The new set of patients is defined through s variables $D' = \{d'_1, \ldots, d'_t\} d'_i \in R^s, s \leq n$.

The second expert, specialized in clustering techniques, uses the information obtained in the previous step to classify the patient into a leukemia group. The patients are first grouped into clusters. The data coming from the pre-processed/filtered phase consists of a group of patients $D' = \{d'_1, \ldots, d'_t\}$ con $d'_i \in R^s, s \leq n$, each one characterized by a set of meaningful attributes $d_i = (x_{i1},\ldots, x_{is})$, where x_{ij} is the fluorescence value of the probe i for the patient j. In order to create clusters and consequently obtain patterns to classify the new patient, the system implements a novel neural network based on the ESOINN [10]. The structure of this neural network has been described in detail in Section 3.2. The network classifies the patients by taking into account their proximity and their density, in such a way that the result provided is a set G where $G = \{g_1,\ldots, g_r\} r < s$. $g_i \subset D$, $g_i \cap g_j = \phi$ with $i \neq j$ and i, $j < r$. The set G is composed of a group of clusters, each of them containing patients with a similar disease. The clusters have been constructed by taking into account the similarity between the patient's meaningful symptoms. Once the clusters have been obtained, the system can classify the new patient by assigning him to one of the clusters. The new patient is defined as d'_{t+1} and his membership to a group is determined by a similarity function defined in (9). The result of the phase is a group of clusters $G = \{g_1, \ldots, g'_i, \ldots, g_r\} r < s$ where $g'_i = g_i \cup \{d'_{t+1}\}$.

Finally, the third expert performs the extraction of knowledge with the CART [11] technique presented in Section 3.3 of this article. The information that is input in this module is made up of the groups selected during the clustering phase, $G = \{g_1,\ldots, g_r\}$, and the influence that each of the genes has on the classification process is calculated. A decision tree is created to generate the rules that determine the influence that each of the genes has in the classification. This then becomes the information that is returned by this module. Specifically, the information that the expert system returns is represented as $g_i = (f_1,\ldots, f_s)$ with $f_i = (z_i, \text{min}, \text{max})$ where z_i is the value of fluorescence, min is the minimum value and max is the maximum value.

4. Results

The Cancer Institute of the University of Salamanca conducts various studies regarding the detection, prediction and treatment of cancer. The department of hematology focuses on the study of blood cancer, or leukemia. This department works with 6 types of pathologies: ALL, AML, CLL, CML, MDS, and NOL (where A is acute, C is chronic, L is lymphocytic, and M is myeloid). Each of these pathologies is characterized by its own symptomatology and by the effects that indicate their development and possible treatments.

- ALL. This is a type of cancer of the blood and bone marrow caused by an abnormal proliferation of lymphocytes.
- AML. This is a type of cancer in the bone marrow characterized by the proliferation of myeloblasts, red blood cells or abnormal platelets.
- CLL. This is a type of cancer characterized by a proliferation of lymphocytes in the bone marrow.
- CML. This is caused by a proliferation of white blood cells in the bone marrow.
- MDS (Myelodysplastic Syndromes). This refers to a group of diseases of the blood and bone marrow in which the bone marrow does not produce a sufficient amount of healthy cells. This can progress to acute leukemia.
- NOL (Normal). No leukemias.

The present case study uses 248 samples obtained by analyses performed on patients either directly from the bone marrow that were hybridized and analyzed with exon arrays manufactured by Affymetrix. The purpose of the tests is to evaluate the validity of the model of experts presented in Section 3 of this article. To this end, the model of experts is applied to the available samples, which results in a classification of patients for each of the groups considered. The classification is compared against the results obtained by the Cancer Institute of the University of Salamanca which employs traditional methods.

In the leukemia studies based on data from exon arrays, the process of filtering data acquires special importance. In the experiments reported in this paper, we worked with a database of bone marrow cases from 248 adult patients with five types of leukemia, plus a group of 16 samples belonging to healthy persons (no leukemias). The data consisted of around 5,500,000 scanned intensities. The system presents a novel technique to reduce the dimensionality of the data. The total number of variables selected in our experiments was reduced to 883, which increased the efficiency of the cluster probe, while the traditional tools, such as the Affymetrix expression array console or Partek Suite, only allowed working with the initial number of variables. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Cancer Institute of the University of Salamanca. The error rates have remained fairly low especially for cases where the number of patients was high. In an attempt to increase the reduction of the dimensionality of the data we applied principal components (PCA) [43], following the method of eigen values over 1. A total of 112 factors were generated, collecting 96% of the variability. However, this reduction of the dimensionality was not appropriate in order to classify patients correctly, so this step was removed from the recovery phase. Fig. 3 shows the classification performed for patients from groups CLL and ALL. The X axe represents the probes used in the classification (883) and the Y axe represents the individuals. As can be seen in Fig. 3a, represented in black, most of the people of the CLL group are together, coinciding with the previous classification given by the experts at the Cancer Institute of the University of Salamanca. Only a small portion of the individuals departed from the initial classification. Fig. 3b shows the classification obtained for the ALL patients. In Fig. 3b it can be seen that, although the ranking is not bad, the proportion of individuals misclassified is higher. Groups that have fewer individuals are those with a higher classification error.

During the filtering phase, we configured several values for the significance level in order to check the possibility of increasing or decreasing the filtering, depending on the selected parameters. Table 3 shows the results obtained after configuring the parameters which affect the significance level, and the misclassification errors obtained from different configurations. The value of the parameter k has been changed in order to study the variability, as well as the value of the parameter α, uniform test, and the value of the Pearson lineal correlation. As can be seen in Table 3, the parameter with a higher influence in the final number of filtered elements is

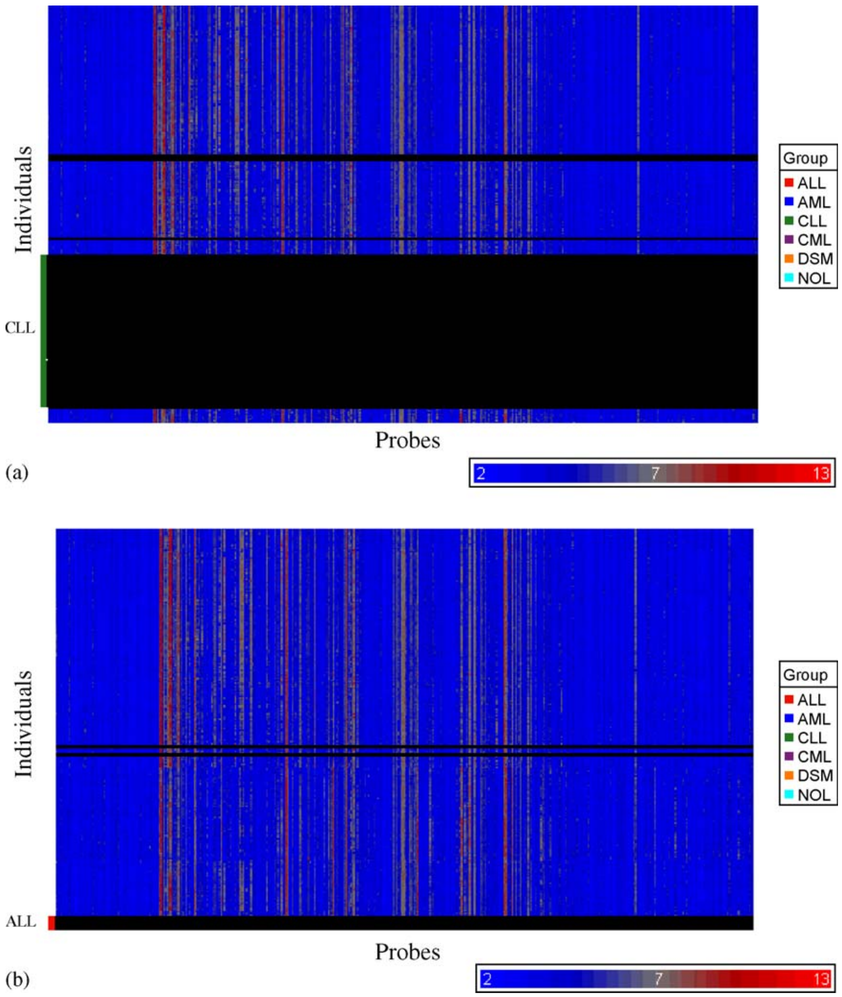


Figure 3 Classification obtained for (a) ALL patients and (b) CLL patients. The *X* axe represents the probes, while the *Y* axe represents the individuals. Each of the values obtained correspond to the fluorescence intensity for an individual and probe. At the bottom of the image it is shown the fluorescence scale of values, the lowest level is two (blue) while the highest is 12 (red) In (a), in the bottom of the figure, it is shown a slide representing those individuals assigned to a group, while the black lines represent the individuals that really belong to that group ALL. In (b) It is represented the same information for the group CLL. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

the one which corresponds to the significance level of the uniform test, while the parameter with a lower influence is the correlation test. Even though for values of α minor to 0.1 we can obtain the higher filtering, we decided to opt for a more conservative posture, choosing a value 0.1 for α. In this sense, the row in bold in Table 3 shows the final configuration for the filtering stage.

In a similar way we proceeded to evaluate the classification for the rest of the groups. Fig. 4a shows the total number of patients with leukemia from each group (11 in ALL, 53 in AML, 95 in CCL, 26 in CML, and 47 in MDS) and the number of misclassifications (4 in ALL, 12 in AML, 4 in CCL, 7 in CML, and 6 in MDS). As can be seen in Fig. 4a, groups with fewer patients are those with a greater error rate. Fig. 4b shows the percentage of error in each group. Once the validity of the method of filtration for selecting the most important variables for classification was verified, the next step in the evaluation was to assess the functioning of the classification process. The system was tested with 15 new patients, who were classified with both the KNN [40] process and the extraction of knowledge technique, following the decision tree shown in Fig. 5, the patients were assigned to the expected groups. Only one of the patients was assigned to a different group by both methods. The healthy patients were eliminated in order to proceed with the classification.

Table 3 Filtering. This table shows the number of probes selected as the confidence levels are modified in the different phases of the filtering stage. The final column represents the classification error obtained after the KNN clustering.

Variability (z)	Uniform (α)	Correlation (α)	Probes	Errors
−1.0	0.25	0.95	2965	42
−1.0	0.15	0.90	1450	
−1.0	0.15	0.95	1488	40
−0.5	0.15	0.90	1368	
−0.5	0.15	0.95	1445	
−1.0	**0.1**	**0.95**	**883**	**41**
−1.0	0.05	0.90	384	
−1.0	0.05	0.95	388	51
−0.5	0.05	0.9	362	
−0.5	0.05	0.95	363	
−1.0	0.01	0.95	66	92

The final classification was compared with the data obtained using a dendogram [44] and partitioning around medoids (PAM) [45]. The proportion of errors in every group was calculated and the Kruskal—Wallis [42] test was applied to determinate if the median of these proportions were equal. The results are shown in Table 4. As previously explained, Table 4 presents the results obtained after applying a non-parametric test which allows comparing equal proportions in misclassification.

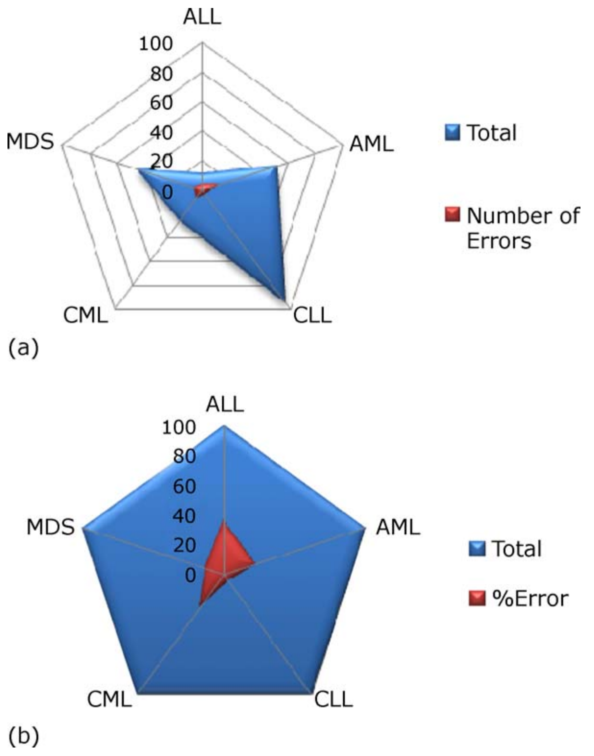


Figure 4 Classification errors (a) numerical (b) percentage.

In this case, taking into account the mentioned patient classification methods, the results show that the proportions of errors is lower if the mixture of experts is used.

Once the individuals have been classified, the extraction of knowledge method is applied by means of the technique explained in Section 3.3. In the first step, the Kruskal—Wallis [42] non-parametric test was applied for equality between medians, where all variables with a median equal for all groups were eliminated. As a result, 35 variables were eliminated, leaving a total of 848. In the next step, the same test was applied on all group pairs, from which we determined that the median among all group pairs was different for each of the variables. Fig. 6 shows the pairs compared for the medians of all groups. The value of the non-parametric contrast of the Kruskal—Wallis test is represented for those probes for which it has been discarded that their equality between medians be the same for all the groups. Each of the rows in Fig. 6 represents each of the possible pairs, and the columns represent the 848 variables. As can be seen in Fig. 6, the medians for almost all of the pairs are different. There are only 3 bands that show values for which the equality of medians can be noted in the variables. Because these cannot be extended to the other pairs, the elimination of waves was considered to be completed. The graphic shown in Fig. 6 also provides an indication of which of the groups are going to be difficult to differentiate. By observing the graphic, we can conclude, for example, that the patients of type AML—MDS that correspond to the second red band will be problematic. By looking at Fig. 5 and Fig. 7, we can prove that the majority of the individuals incorrectly classified as type AML could be classified as MDS.

After completing the previous steps, the extraction of knowledge method was applied to the resulting groups. The groups provided by the ESOINN network are considered together with the information (cases) of previous classifications.

Table 4 Comparison of clustering methods. The table shows a comparison of equality of medians for the different clustering methods used. The variable considered is the number of classification errors in each of the groups. The comparison is based on the Kruskal—Wallis non-parametric test. Comparison of methods. * different median and = equal, (−) median of column less than median of row.

	Expert	Dendogram	PAM
Expert			
Dendogram	*(−)		
PAM	*(−)	*(−)	

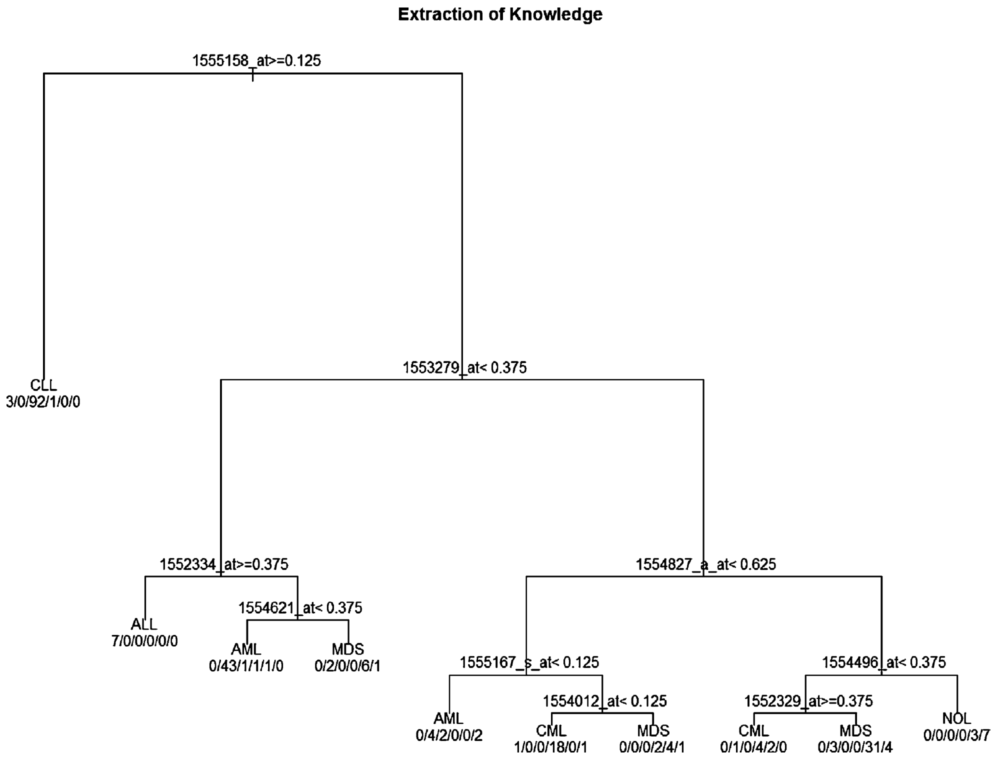


Figure 5 Decision tree for classifying patients. The values of the leaf nodes represent the predicted group and the number of elements assigned to each of the groups following the order (ALL, AML, CLL, CML, NOL, MDS). The rest of the nodes represent the probe and the fuzzy value to compare the individual to classify. If the condition is true, then the branch on the left is selected, otherwise, the branch on the right is selected. The tree helps to obtain an explanation of the reason why an individual has been assigned to a group.

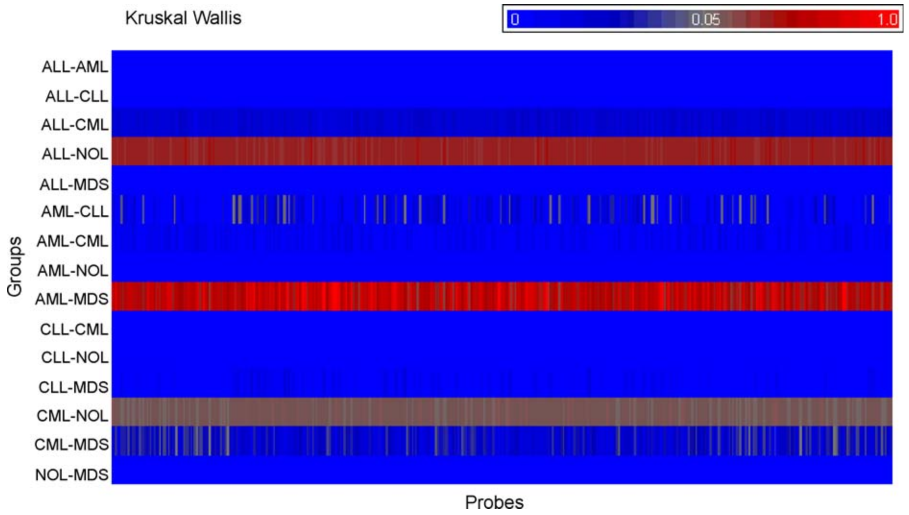
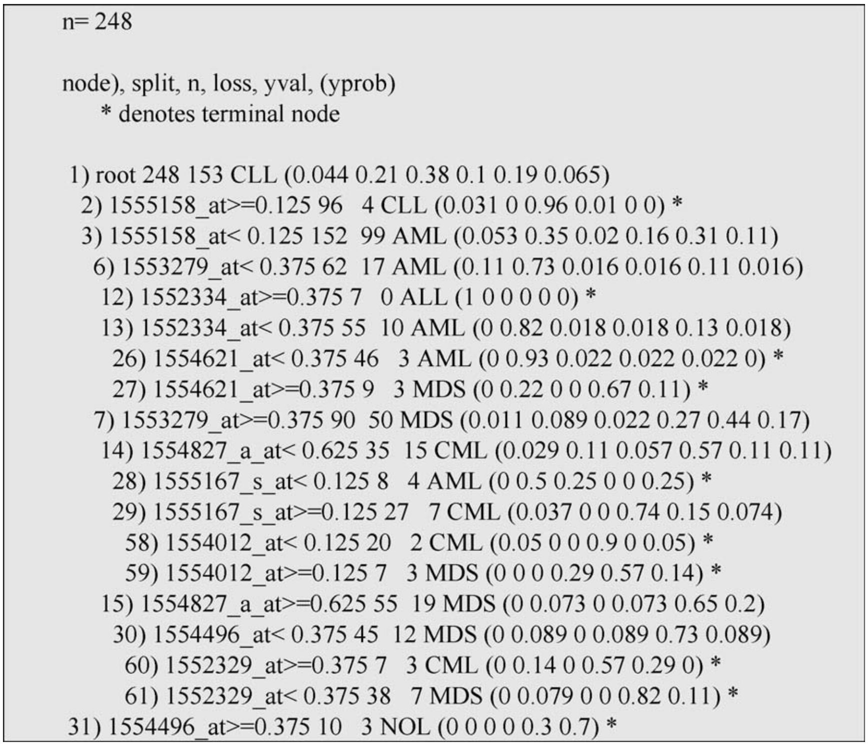


Figure 6 Test of equality for the medians of all group pairs. The *X* axe represents the probes and the *Y* axe represents the pairs of groups. The color represents the value obtained in the Kruskal—Wallis hypothesis contrast. The red bands represent those groups which probes contain similar values. The scale of colors represents the values of the test. The blue region [0, 0.05] is consistent with values of the test for which H_0 is rejected while the red color corresponds to the region which H_0 is accepted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)



```
n= 248

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 248 153 CLL (0.044 0.21 0.38 0.1 0.19 0.065)
   2) 1555158_at>=0.125 96   4 CLL (0.031 0 0.96 0.01 0 0) *
   3) 1555158_at< 0.125 152  99 AML (0.053 0.35 0.02 0.16 0.31 0.11)
     6) 1553279_at< 0.375 62  17 AML (0.11 0.73 0.016 0.016 0.11 0.016)
      12) 1552334_at>=0.375 7   0 ALL (1 0 0 0 0 0) *
      13) 1552334_at< 0.375 55  10 AML (0 0.82 0.018 0.018 0.13 0.018)
        26) 1554621_at< 0.375 46   3 AML (0 0.93 0.022 0.022 0.022 0) *
        27) 1554621_at>=0.375 9   3 MDS (0 0.22 0 0 0.67 0.11) *
     7) 1553279_at>=0.375 90  50 MDS (0.011 0.089 0.022 0.27 0.44 0.17)
      14) 1554827_a_at< 0.625 35  15 CML (0.029 0.11 0.057 0.57 0.11 0.11)
        28) 1555167_s_at< 0.125 8   4 AML (0 0.5 0.25 0 0 0.25) *
        29) 1555167_s_at>=0.125 27   7 CML (0.037 0 0 0.74 0.15 0.074)
          58) 1554012_at< 0.125 20   2 CML (0.05 0 0 0.9 0 0.05) *
          59) 1554012_at>=0.125 7   3 MDS (0 0 0 0.29 0.57 0.14) *
      15) 1554827_a_at>=0.625 55  19 MDS (0 0.073 0 0.073 0.65 0.2)
        30) 1554496_at< 0.375 45  12 MDS (0 0.089 0 0.089 0.73 0.089)
          60) 1552329_at>=0.375 7   3 CML (0 0.14 0 0.57 0.29 0) *
          61) 1552329_at< 0.375 38   7 MDS (0 0.079 0 0 0.82 0.11) *
 31) 1554496_at>=0.375 10   3 NOL (0 0 0 0 0.3 0.7) *
```

Figure 7 Detailed information from the decision tree. Each of the rows corresponds to a tree branch that contains the number of assigned nodes, the condition, the number of nodes correctly classified, the misclassified nodes, and the class they belong to. The probability of each of the classes being assigned to the nodes (ALL, AML, CLL, CML, NOL, MDS) is indicated in parentheses.

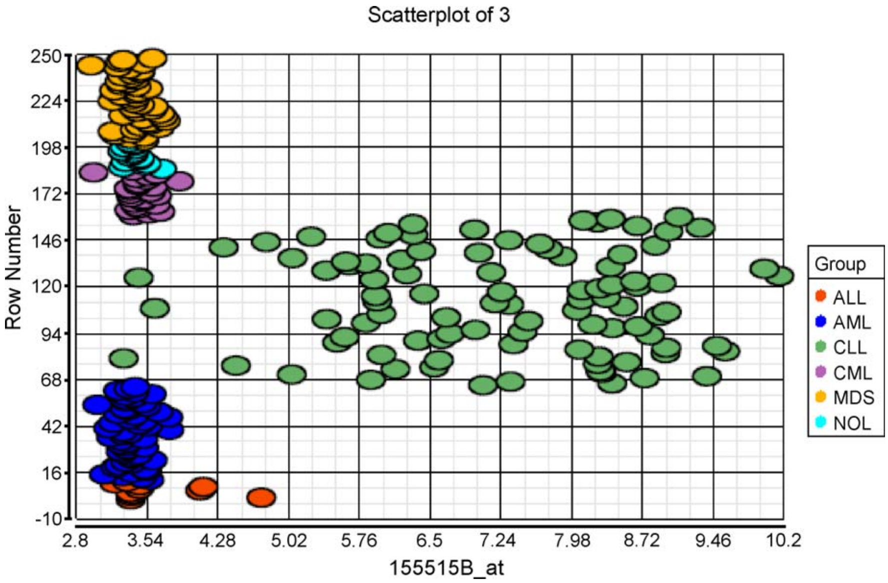


Figure 8 Intensity of 1555158_at. Shows the values of the probe located in the root of the decision tree created by means of CART. As can be seen, the probe allows obtaining a good classification for the individuals belonging to the CLL group.

The network is only used for individuals without previous classification. If the network is used with an individual previously classified and the result differs from the initial one, then it is necessary the opinion of a human expert. In order to perform the extraction of knowledge, as indicated in Section 3.3, a fuzzy logic discretization process was applied to the values. The results obtained were very satisfactory since with only a few waves it is possible to classify patients simply and efficiently. Fig. 5 shows the decision tree that was obtained. The top of each branch displays the decision wave, while the tree leaves display the classification group and the number of elements classified for each type (ALL, AML, CLL, CML, NOL, MDS). Therefore, in order to classify an individual, one would simply start at the top of the tree and verify the value of the waves. Figs. 8 and 9, show a graphical representation of the tree generated. Fig. 8 shows the values of the probe root 1555158_at. It is possible to observe that the individuals belonging to the CLL group possess values which are superior to those of other individuals, and this fact allow us to easily identify the individuals of CLL class. Fig. 9

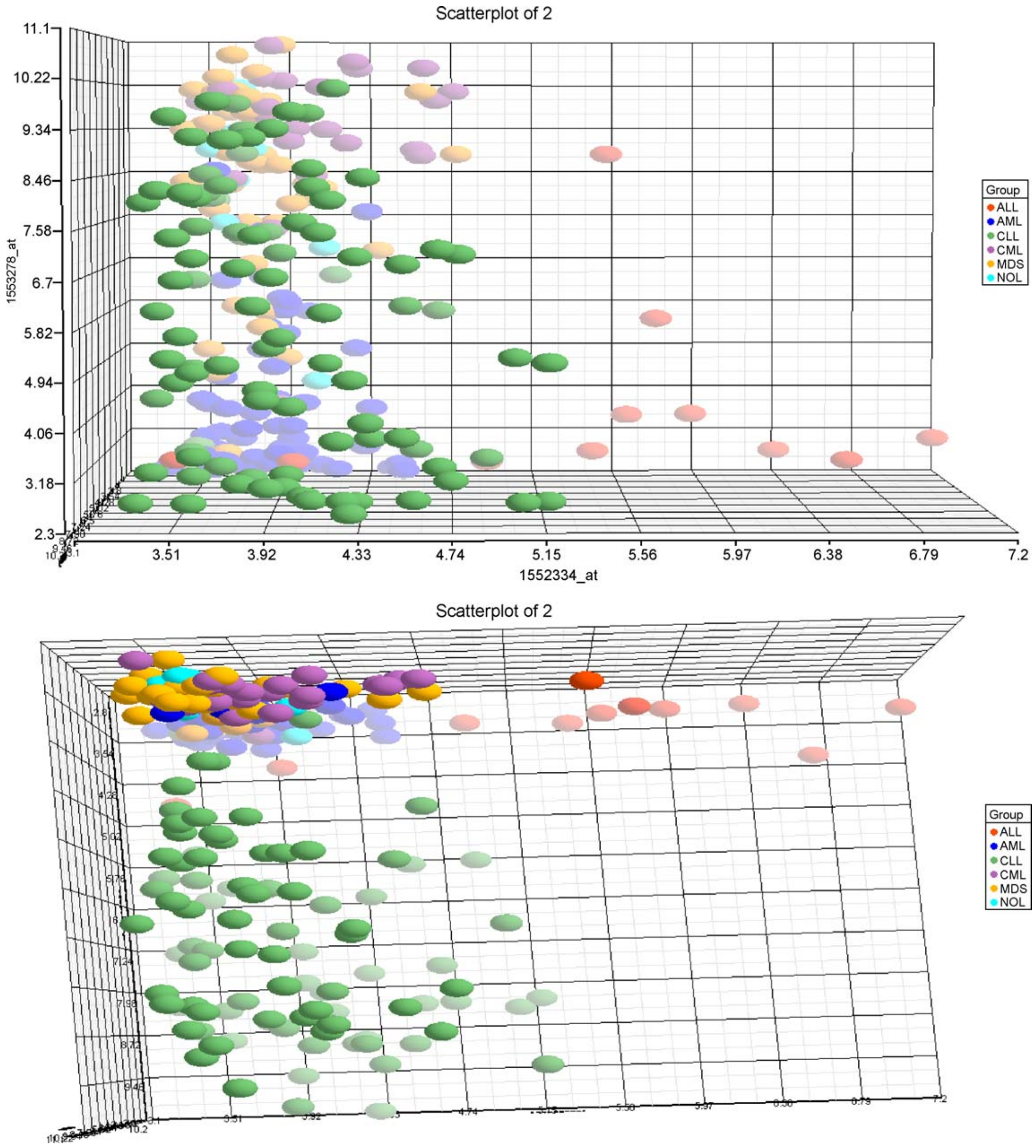


Figure 9 Classification CLL—ALL. Representation of the probes of the decision tree which classify the CLL and ALL to 1555158_at, 1553279_at and 1552334_at. Both images represent the same information, but from different angles.

shows a representation of the first 3 probes of the tree, 1555158_at, 1553279_at and 1552334_at. These probes allow us to separate the individuals from groups CLL and ALL with a high success rate. Both figures represent the same information, but from different perspectives.

Fig. 7 shows the information from the decision tree in detail. Each of the rows corresponds to a tree branch that contains the number of assigned nodes, the condition, the number of nodes correctly classified, the misclassified nodes, and the class they belong to. The probability of each of the classes being assigned to the nodes (ALL, AML, CLL, CML, NOL, MDS) is indicated in parentheses. Fig. 10 is a graphical representation of the classification errors obtained from the decision tree. Fig. 10a shows the number of erroneous and successful classifications in each group. Fig. 10b shows the same data in percentages. As can be seen in Fig. 10, the classification errors were hardly significant.

Because the system was intended to classify leukemia patients, those not suffering from this disease (from group NOL) were eliminated, and the same test was conducted again. Fig. 11 shows the decision tree obtained from this classification. As can be seen, the complexity of the tree has been reduced from 10 leaves to 8. However, note that discriminating waves have not varied with respect to those shown in Fig. 5. Additionally, the final percentage of error for all the groups remains constant.

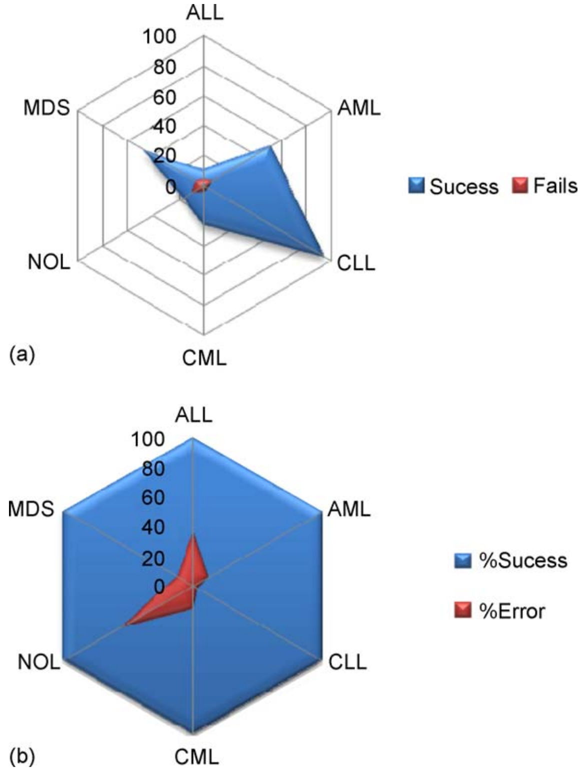


Figure 10 Classification errors (a) numerical (b) percentage using the decision tree.

Once the classification model was obtained, its validity was confirmed. In order to accomplish this, a detailed prediction was made from the samples corresponding to individuals from the Cancer Institute database by determining the probability of assigning each individual to each one of the groups. This made it possible to easily observe which individuals were misclassified and the degree of certainty for each classification based on probability. Fig. 12 shows the different probabilities for assigning each individual to each of the groups according to the CART algorithm [11]. The x-axis represents the individuals and the y-axis the probability of assignment to each of the groups. The line at the bottom of the graph in Fig. 12 corresponds to the origin of each individual, while the top part of the graph shows the group to which the patient is finally assigned. This probability is calculated taking into account the possible patient assignation related to the selected probes. The CART algorithm allows obtaining the probes of more influence for the classification of each of the types. Fig. 13 shows a graphical representation of the three probes that better classify the patients with leukemia of CLL class. As can be seen in Fig. 13a and b, any of the axes allows a separation of the CLL individuals from the rest in an efficient manner.

One of the great contributions of the model presented is the ability to work with data from exon arrays because of its great capacity for selecting significant variables. Nowadays very few tools are capable of working with data of this kind, due to its high dimensionality. The proposed model resolves this problem by using a technique that detects the most important genes for the classification of diseases by analyzing the available data. As demonstrated, the proposed system allows the reduction of 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 use of ESOINN 0 neural networks and a technique for discovering CART rules [11] that can be viewed as general knowledge summarizing the relevance of the acquired knowledge. The results obtained from empirical studies are promising and highly appreciated by laboratory specialists, as they are provided with a tool that allows the detection of genes and those variables that are most important for discovering a pathology, and facilitates a reliable classification and diagnosis, as shown by the results presented in this paper.

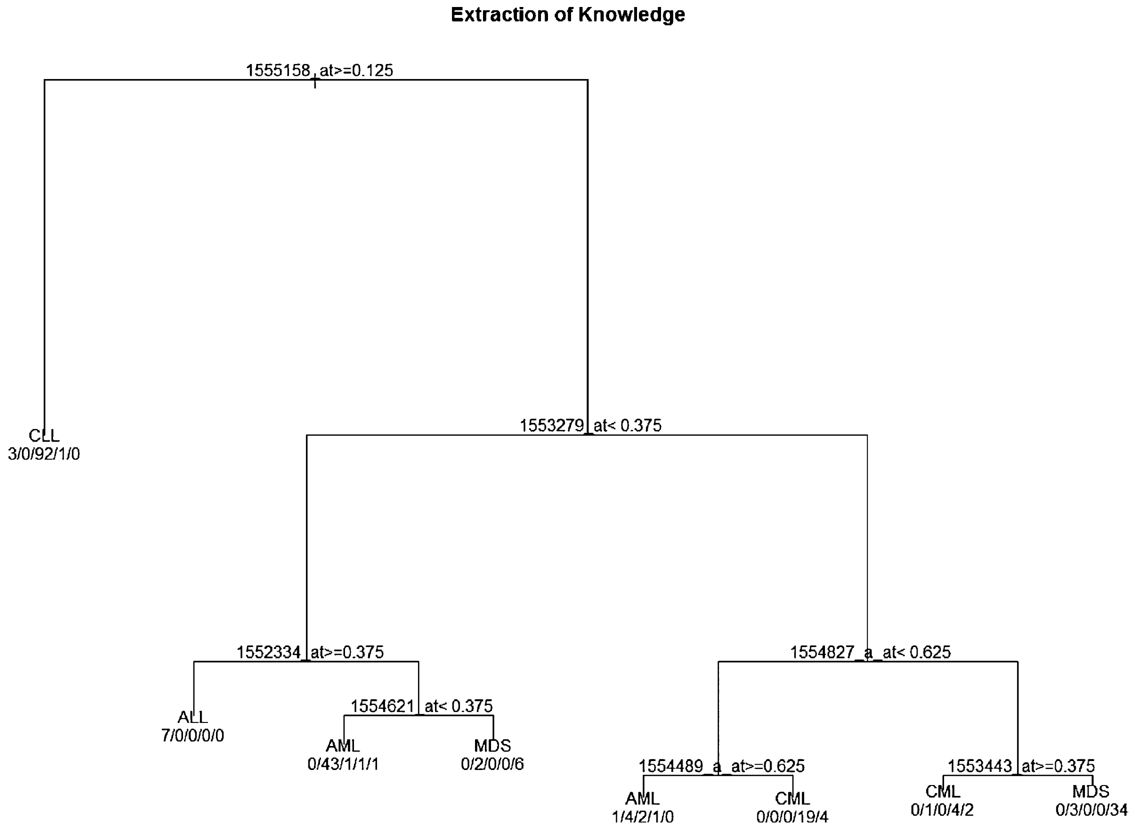


Figure 11 Decision tree for the classification of leukemia patients with out NOL. The values of the leaf nodes represent the predicted group and the number of elements assigned to each of the groups following the order (ALL, AML, CLL, CML, MDS). The rest of the nodes represent the probe and the fuzzy value to compare the individual we are trying to classify. If the condition is true, then the branch on the left is selected, otherwise, the branch on the right. In this way it is possible to obtain an explanation of the reason why an individual has been assigned to a group.

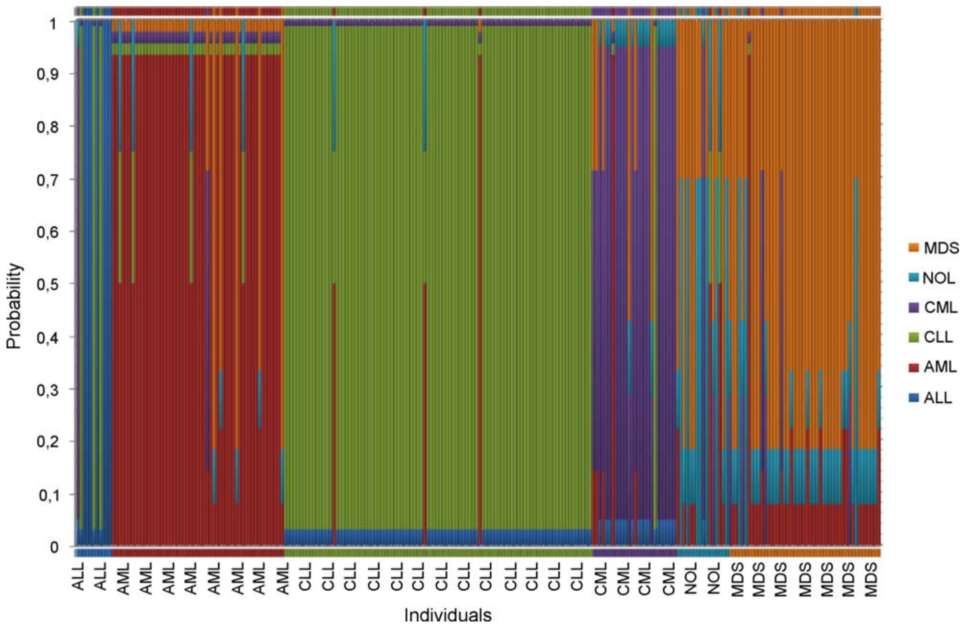


Figure 12 Probability prediction for assigning each individual to each of the groups, according to the CART algorithm.

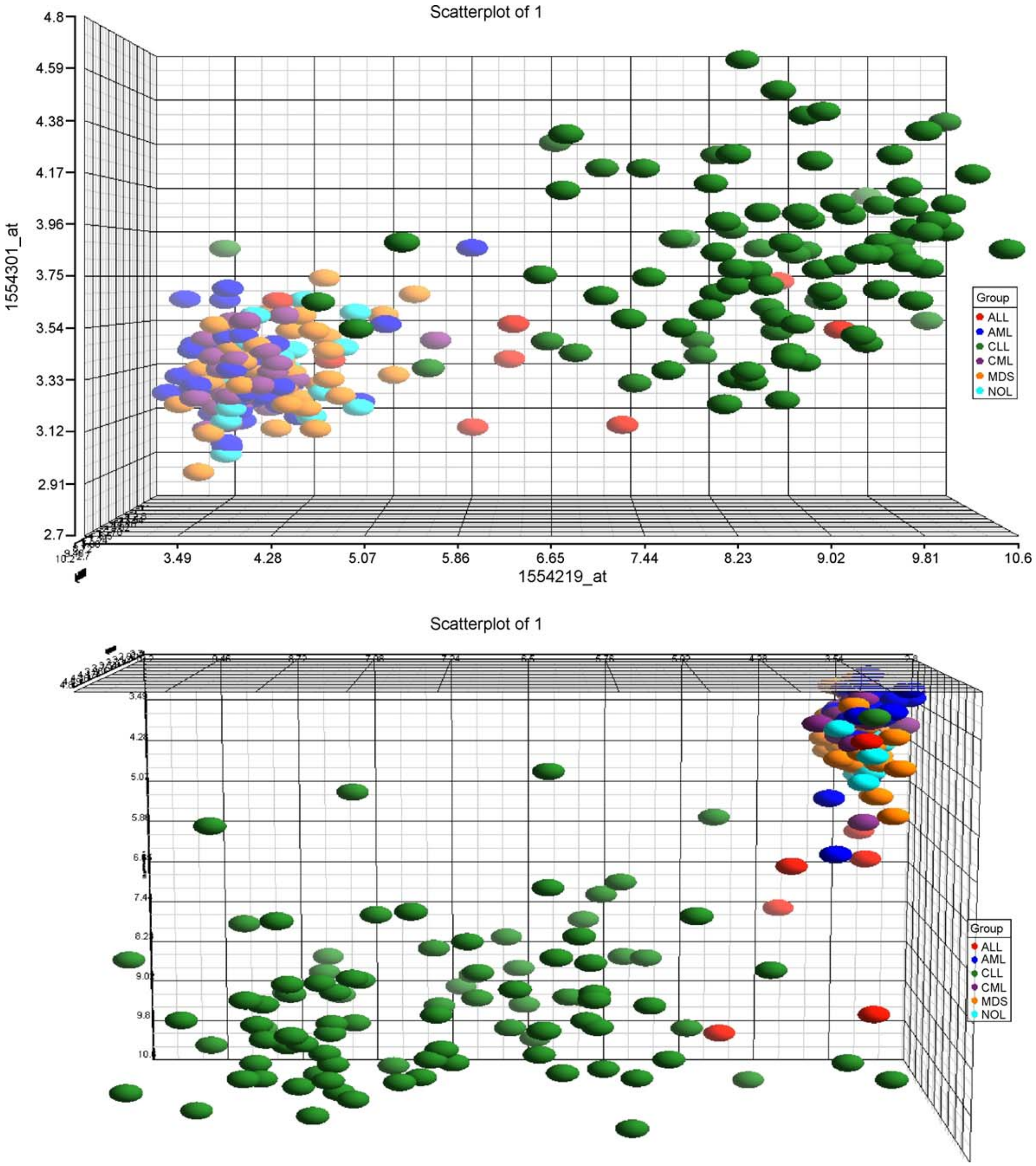


Figure 13 Classification CLL from the most important probes extracted by the CART algorithm. Each of the axes represents one of these probes extracted by CART for the classification of the CLL group.

5. Conclusions

This study has presented a model of experts that uses exon arrays [4] to perform an automatic diagnosis of cancer patients. It is a system that incorporates experts at each phase of the microarray analysis, a process that is capable of extracting knowledge from diagnoses that have already been performed, and that has been used to increase the efficiency of new diagnoses. The model combines novel techniques that reduce the dimensionality of the original set of data under study, pre-processing and data filtering techniques, a novel method of clustering for classifying patients, and modern extraction of knowledge techniques. The system works in a way that is similar to how human specialists work in the laboratory, but is also capable of working with large amounts of data and making decisions automatically, thus significantly reducing both the time needed for making predictions, and the rate of human error. The study presented within the scope of this research focused on identifying the main variables for each disease so that patients can be classified accordingly. It would be interesting to conduct future studies to analyze if different classifications exist depending on whether the samples are obtained from the bone marrow or blood.

The advantage of using a mixture of experts lies in the flexibility and adaptability it affords to the needs of the problem being studied. Furthermore, it facilitates the incorporation of different experts in each of the phases of the model, which can offer different perspectives for approaching the problem at each particular phase. These perspectives can be compared so that the optimal decision can be selected for each specific situation. One of the greatest contributions of the model presented is the ability it has to work with exon array data 0. Nowadays, very few tools are capable of working with this type of data because of the high dimensionality. The proposed model resolves this problem by using a technique that detects the importance of the genes for the classification of the diseases by analyzing the available data.

For the time being, three experts have been designed, one for each phase of the model. They can adapt to the needs of the problem of diagnosing leukemia patients and present novel characteristics. We have proposed a system to reduce the dimensionality based on the application of new filtering techniques. Additionally, we have presented a clustering technique based on the use of ESOINN neural networks 0 which allow the classification of individuals. Finally, we have developed an improved version of the CART 0 extraction of knowledge algorithm. The results obtained from empirical studies are promising and much appreciated by laboratory specialists, as they are provided with a tool that allows the detection of the most important genes and variables needed for discovering a pathology, and facilitates a reliable classification and diagnosis, as shown by the results presented in this study. Traditional filtering techniques are inefficient in terms of selection of important probes for the classification process, which greatly complicates the extraction of knowledge. The medical staff of the Cancer Institute appreciates a system capable of providing automatic filtering and classification, but particularly praises the facility of decision support, as it helps to explain the analysis process by means of rules, and provides valuable knowledge on the meaningful probes that allow classifying the individuals.

Acknowledgments

This work was supported in part by the projects MEC THOMAS TIN2006-14630-C03-03 and JCYL SA071A08. We will like to thank also our colleges from the Cancer Institute of the University of Salamanca.

References

[1] Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine. New York: Springer; 2006.

[2] Tsoka S, Ouzounis C. Recent developments and future directions in computational genomics. FEBS Letters (Elsevier Amsterdam) 2000;480(1):42—8.

[3] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860—921.

[4] Affymetrix, GeneChip Human Exon 1.0 ST Array, http://www.affymetrix.com/products/arrays/specific/Exon.affx [accessed: 1 June 2008].

[5] Rubnitz JE, Hijiya N, Zhou Y, Hancock ML, Rivera GK, Pui C. Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatric Blood & Cancer 2005;44(2):138—41.

[6] Affymetrix, Exon Array Design Datasheet, http://www.affymetrix.com/support/technical/datasheets/Exon_arraydesign_datasheet.pdf; 2008 [accessed: 1 June 2008].

[7] Armstrong NJ, Van de Wiel MA. Microarray data analysis: from hypotheses to conclusions using gene expression data. Cellular Oncology 2004;44(2):279—90.

[8] Quackenbush J. Computational analysis of microarray data. Nature Review Genetics 2001;2(6):418—27.

[9] Zweiger G. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. Trends in Biotechnology 1999;17(11):429—36.

[10] Furao S, Ogura T, Hasegawa O. An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks 2007;20:893—903.

[11] Breiman L, Fried man JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, California: Wadsworth International Group; 1984.

[12] Kearney L. Molecular cytogenetics. Best Practice & Research Clinical Haematology 2001. 645—U3.

[13] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403: 503—11.

[14] Schena M, Shalon D, Davis R, Brown PO. Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 1995;270:467—70.

[15] The Leukemia & Lymphoma Society, http://www.leukemia-lymphoma.org/hm_lls; 2008 [accessed: 1 June 2008].

[16] Fabiola G. Statistic Data on Leukemia, Ezine Articles, http://ezinearticles.com/?Statistic-Data-on-Leukemia&id=465391; 2008 [accessed: 1 June 2008].

[17] SEER (Surveillance Epidemiology and End Results), U.S. National Cancer Institute, http://seer.cancer.gov/; 2007 [accessed: 1 June 2008].

[18] Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DH. High density synthetic oligonucleotide arrays. Nature Genetics 1999;21:20—4.

[19] Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter 2003;5(2):1—5.

[20] Affymetrix Statistical Algorithms Description Document, http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf; 2002 [accessed: 1 June 2008].

[21] Affymetrix, Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf; 2005 [accessed: 1 June 2008].

[22] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249—64.

[23] Kohonen T. Self-organized formation of topologically correct feature maps. Biological cybernetics, vol. 43. Berlin: Springer; 1982. p. 59—69.

[24] Fritzke B. A growing neural gas network learns topologies. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems, vol. 7. Cambridge, MA; 1995. p. 625—32.

[25] Martinetz T. Competitive Hebbian learning rule forms perfectly topology preserving maps. In: Gielen S, Kappen B, editors. ICANN'93: international conference on artificial neural networks. Holland, Amsterdam: Springer; 1993. p. 427—34.

[26] Martinetz T, Schulten K. A neural-gas network learns topologies. In: Kohonen T, Makisara K, Simula O, Kangas J, editors. Artificial neural networks. Amsterdam: Elsevier; 1991. p. 397—402.

[27] Shen F. An algorithm for incremental unsupervised learning and topology representation. Ph.D. thesis. Tokyo Institute of Technology; 2006.

[28] Mena J. Data mining your website. New Cork: Digital Press; 1999.

[29] Pawlak Z. Rough sets. International Journal of Information & Computer Sciences 1982;11:341—56.

[30] Hunt EB, Marin J, Stone PJ. Experiments in induction. New York: Academic Press; 1966.

[31] Quinlan JR. Discovering rules by induction from large collections of examples. In: Michie D, editor. Expert systems in the micro electronic age. Edinburgh: Edinburgh University Press; 1979. p. 168—201.

[32] Murthy SK, Kasif S, Salzberg S. A system for the induction of oblique decision trees. Journal of Artificial Intelligence Research 1994;2:1—33.

[33] Cestnik B, Kononenko I, Bratko I. ASSISTANT 86: a knowledge elicitation tool for sophisticated users. In: Bratko I, Lavrac N, editors. Progress in machine learning. Wilmslow, England: Sigma Press; 1987. p. 31—45.

[34] Quinlan JR. C4.5: programs for machine learning. San Francisco, California, USA: Morgan Kaufmann Publishers Inc.; 1993

[35] Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine 2007;37(2):251—61.

[36] Steinberg D, Colla P. CART: tree-structured non parametric data analysis. San Diego, California, USA: Salford Systems; 1995.

[38] Brunelli R. Histogram analysis for image retrieval. Pattern Recognition 2001;34:1625—37.

[39] Corchado JM, Bajo J, De Paz Y, De Paz JF. Integrating case planning and RPTW neuronal networks. Expert Systems with Applications 2009;43(2). doi: 10.1016/j.eswa.2008.07.029.

[40] Fix E, Hodges JL. Discriminatory analysis, nonparametric discrimination consistency properties. Technical Report 4. United States Air Force, Randolph Field, TX; 1977.

[41] Gariepy R, Pepe WD. On the level sets of a distance function in a Minkowski space. Proceedings of the American Mathematical Society 1972;31(1):255—9.

[42] Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. Journal of American Statistics Association 1952;47 (260):583—621.

[43] Jolliffe I. Principal component analysis. Springer series in statistics, Second edition, Berlín: Springer; 2002.

[44] Saitou N, Nie M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 1987;4:406—25.

[45] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley series in probability and statistics. New York: John Wiley and Sons Ltd.; 1990.