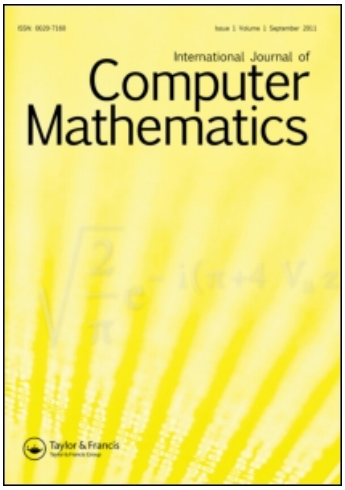

International Journal of Computer Mathematics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713455451

An adaptive algorithm for feature selection in pattern recognition

Juan F. De Paz[a]; Sara Rodríguez[a]; Vivian F. López[a]; Javier Bajo[a]
[a] Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, Spain

First published on: 27 January 2011

To cite this Article De Paz, Juan F. , Rodríguez, Sara , López, Vivian F. and Bajo, Javier(2011) 'An adaptive algorithm for feature selection in pattern recognition', International Journal of Computer Mathematics, 88: 9, 1932 — 1940, First published on: 27 January 2011 (iFirst)
To link to this Article: DOI: 10.1080/00207160.2010.484100
URL: http://dx.doi.org/10.1080/00207160.2010.484100

International Journal of Computer Mathematics
Vol. 88, No. 9, June 2011, 1932–1940

Taylor & Francis
Taylor & Francis Group

An adaptive algorithm for feature selection in pattern recognition

Juan F. De Paz, Sara Rodríguez, Vivian F. López and Javier Bajo*

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

(*Received 30 September 2009; revised version received 20 January 2010; accepted 26 March 2010*)

With the most recent advances in bioinformatics, the amount of information available for analysing certain diseases has increased considerably. Specifically, the use of microarrays makes it possible to obtain information on genetic patterns. The analysis of this information requires the use of new computational models and the modification of existing models so that it becomes possible to work with such an elevated amount of data. This study will demonstrate the integration of an expression analysis in a case-based reasoning system that can apply data mining techniques to classify and obtain patterns that have been stored in a case database for leukaemia patients.

Keywords: case-based reasoning; SODTNN; leukaemia classification; decision tree; problem solving; logic in artificial intelligence; distributed artificial intelligence

2000 AMS Subject Classifications: 68T20; 68T27
ACM Computing Classification System Code: I.2.11

1. Introduction

The last few years have seen great advances in the fields of bioinformatics and biomedicine. The incorporation of data mining and artificial intelligence into biomedical data has led to significant progress in the prevention and detection of diseases. Genomics is an area of bioinformatics that is in the height of its development and in which the application of techniques is essential to facilitate both the automated treatment of data and extraction of knowledge. There are different fields of study within genomics. One of them is transcriptomics, which involves the study of ribonucleic acid using techniques such as expression analysis. Expression analysis consists of hybridizing a sample from a case study, from which different levels of luminescence are obtained that can be analysed and represented as a data array. The purpose of this article is to study and extract information from biomedical databases that characterizes different individuals based on the luminescence level of microarrays.

DNA microarray technology makes it possible to substantially increase investigations in molecular biology [16,17]. The study of expression analysis is a very important field of scientific

*Corresponding author. Email: jbajope@usal.es

ISSN 0020-7160 print/ISSN 1029-0265 online

DOI: 10.1080/00207160.2010.484100
http://www.informaworld.com

investigations because changes or disease in certain organisms can be a reflection of changes in expression patterns. Expression arrays [17] are a type of microarray that can be used in different approximations for identifying genes that characterize specific diseases [16,17]. An expression analysis basically consists of three phases: standardization and filtering, clustering and classification, and knowledge extraction. These steps are carried out starting with the data from the luminescence values, which are obtained from the probes found within the microarrays. The probes contain a sequence of oligonucleotides that fuses with the DNA-dyed sample and generates luminescence. The amount of information that is provided amounts to several million bits of data for each sample. The current problem deals specifically with the Affymetrix HG U133 Plus 2.0 Array [1], which includes 1,300,000 values for each test. These values correspond to a series of oligonucleotides that are grouped in probes. Each of the chips contains approximately 54,000 probes.

This research will present a computational model that integrates the different phases of expression analysis using a case-based reasoning (CBR) system [10]. Subsequent iterations will be based on the previous predictions, thus incorporating a final learning mechanism. CBR solves problems by using a reasoning cycle comprising four sequential phases: retrieve, re-use, revise, and retain. This process is very similar to the steps carried out in the expression analysis. Specific algorithms are introduced within each of these phases as needed to finalize the expression analysis. During each of the phases, various statistical techniques are integrated for reducing the dimensionality of the data. Clustering and classification techniques are integrated for associating individuals with different classes, and knowledge extraction techniques are used for explaining the final groupings.

The paper is structured as follows. Section 2 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 leukaemia patients. Section 4 presents the results, and finally, conclusions are drawn in Section 5.

2. CBR system for classifying microarray data

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 that describes the initial problem, a solution that provides the sequence of actions carried out in order to solve the problem, and the final state that 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, re-use, revise, and retain.

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 six subphases: pre-processing through robust multi-array average (RMA), removal of control probes, erroneous probes, low variability, uniform distribution, and correlated variables. In the re-use phase, the patients are grouped by means of a Self-Organized Dynamic Tree Neural Network (SODTNN) [5]. Then, the patients without prior classification are assigned to a group. In the revise phase, the classification and regression tree (CART) [2] technique is applied for extracting knowledge about the most important probes for the classification. CART was used because it allows representing knowledge as a tree, thus simplifying the classification process. Finally, in the retain phase, the knowledge is updated.

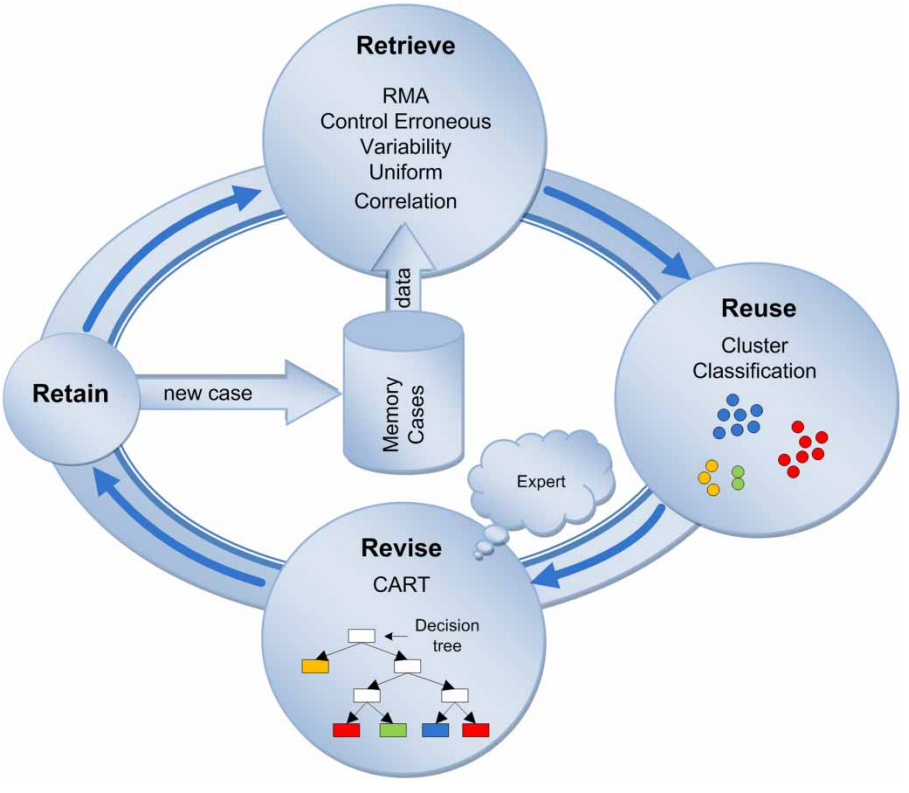

Figure 1. CBR system.

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, this stage is focused on the retrieval of relevant probes. We have incorporated an innovative strategy in which variables are retrieved at this stage and then, depending on the identified variables, the rest of the stages of the CBR are carried out. First, a pre-processing of the data is conducted using RMA. Then, the five filtering subphases 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 subphases are outlined in the following sections.

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 analysing 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 three subphases: background correction, standardization quantile normalization, and summarization. The RMA [7] algorithm is frequently

used for pre-processing Affymetrix microarray data and consists of three steps: background correction using the $\log_2$ of the level of luminescence, standardization based on quantile normalization, and summarization through lineal models. A complete description of RMA and a comparison to related techniques can be seen in [7].

2.1.2 *Control and erroneous*

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 that should contain the same values for all individuals. If they have different values, the case should be discarded. 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. Each of the probes is replicated in different zones of the chip, and if the replicas contain significant differences, then the probe is identified as erroneous.

2.1.3 *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

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

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

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

   where $\mu = 1/N \sum_{j=1}^{N} \sigma_{\cdot j}$ and $\sigma_{\cdot j} = +\sqrt{1/N \sum_{j=1}^{N} (\bar{\mu}_{\cdot j} - x_{ij})^2}$, in which $z_i \equiv N(0, 1)$.
3. Discard probes for which the value of z meets the following condition: $z < -1.0$.

2.1.4 *Uniform distribution*

Finally, all remaining variables that follow a uniform distribution are eliminated. The variables that follow a uniform distribution will not allow the separation of individuals. Therefore, the variables that do not follow this distribution will be really useful variables in the classification of the cases. The contrast of assumptions is explained below, using the Kolmogorov–Smirnov [3] test as an example: H0, the data follow a uniform distribution and H1, the analysed data do not follow a uniform distribution. Statistical contrast:

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

where $D^+ = \max_{1 \le i \le n}\{i/n - F_0(x_i)\}$ and $D^- = \max_{1 \le i \le n}\{F_0(x_i) - (i-1)/n\}$, 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 with the next value:

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

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

2.1.5 *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.

$$r_{x_{\cdot i} y_{\cdot j}} > \alpha, \tag{5}$$

given $\alpha = 0.95$, $r_{x_{\cdot i} y_{\cdot j}} = \sigma_{x_{\cdot i} x_{\cdot j}} / \sigma_{x_{\cdot i}} \sigma_{x_{\cdot j}}$, $\sigma_{x_{\cdot i} x_{\cdot j}} = 1/N \sum_{s=1}^{N} (\bar{\mu}_{\cdot i} - x_{si})(\bar{u}_{\cdot j} - x_{sj})$, where $\sigma_{x_{\cdot i} x_{\cdot j}}$ is the covariance between probes i and j.

2.2 *Re-use*

In the re-use phase, a new patient is assigned to a concrete cluster taking into account the information provided by the probes selected in the retrieve phase. This phase has been divided into two sequential stages: (i) clustering and (ii) classification. The most common techniques to carry out the functionalities required in this phase are dendograms [15], partition around medoids (PAM) [8], or k-means [14]. The main advantage provided by these methods is the ability to deal with a great number of variables, without performing a previous dimensionality reduction. However, these methods are not able to automatically adapt their behaviour to any data distribution, and it is necessary to explicitly fix the number of clusters that will be created. Moreover, they present problems when atypical data are used [5]. As an alternative to these methods, some artificial neural networks are proposed. That is, the case of the self-organizing map (SOM) [9], growing neural gas [9], growing cell structure (GCS) [11], or enhanced self-organizing incremental neuronal network (ESOINN) [6]. All these artificial neural networks have a greater ability to adapt to the data distribution, but present the inconvenience that they require an adaptation of the neurons to the data surface, and in consequence, it is necessary to create new algorithms for facilitating incorporate/remove neurons of the network structure. In previous works, we developed the SODTNN [5] that simplifies the expansion and division processes of the neural network. Figure 2 shows an example of comparison of clusters created using the SOTDTNN and the rest of previously commented methods using R libraries. Each of the colours represents how the elements were assigned to a cluster. A more detailed study can be seen in [5].

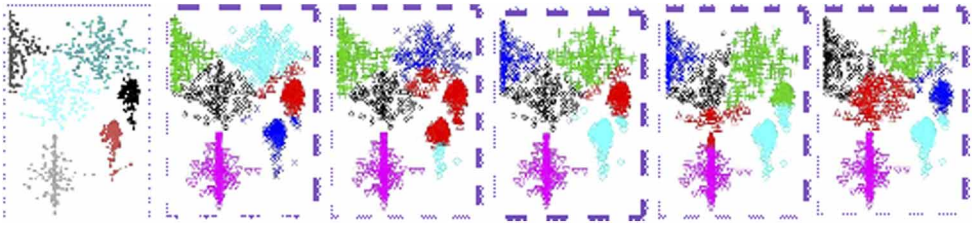

Figure 2. Sequence of clusters generated by SODTNN, PAM, dendrograms, fanny, Diana, Clara in this order.

In summary, the SODTNN uses the connections of the minimal tree constructed using the Kruskal algorithm to detect the number of existing classes and to create clusters. The main difference with neural networks as ESOINN or GCS is that no distinction is made between the input data and the neurons. Thus, the neurons correspond, at the beginning of the training stage, to the position of each of the input data. The learning process is executed iteratively by selecting the neurons for the neural network and updating their position and the position of their neighbours in each iteration. The neurons from the network that define the clusters are updated periodically in a way similar to the Kohonen SOM. By updating automatically, the positions and connections of the neurons can be readjusted in order to complete the division of the clusters. The network randomly selects an initial neuron and brings neighbouring neurons closer in. The neurons are updated according to the hierarchy of the tree. The nomenclature defines T as the set of neurons to be classified and A as the minimum spanning tree that contains all of the nodes from T. The magnitude and the direction of the vector depend on the distance and neighbourhood as indicated in the following algorithm:

1. given $k \in T$ with $a_k \in A$ being the selected neuron, set the value of the neighbouring radius r;
2. begin $i = 1$, $a_s = a_k$;
3. calculate the parent node from the current node $a_t = f^p(a_s)$ and obtain all the sons from a_t, which are defined as $A^i_{a_t}$;
4. for each instance $a_j \in A^i_{a_t}$, update the coordinates for the neuron by following the equation for SOMs;
5. $x_j(t+1) = x_j(t) + \eta(t) \cdot g(i,t) \cdot (x_s(t) - x_j(t))$;
6. where $g(i,t)$ represents the neighbouring function $\eta(t)$, the learning rate [4]

$$g(i,t) = \mathrm{Exp}\left[-\frac{i}{N} \frac{\sqrt{(x_{j1} - x_{s1})^2 + \ldots + (x_{jn} - n_{sn})^2}}{\mathrm{Max}\{d_{ij}\}_{i,j}} - \lambda \frac{i \cdot t}{\beta N} \right],$$

$$\eta(t) = \mathrm{Exp}\left[-\sqrt[4]{\frac{t}{\beta N}} \right];$$

7. where t is the iteration, N the number of elements from group $\#A$, n the dimension of the coordinates, x_{ij} coordinate j for the neuron $i \in T$, with $a_i \in A$, and λ andβ the constants established for 1 and 5, respectively;
8. if $i < r$, set $a_s = a_t$ and increase i;
9. use the same procedure to update the descendents $a_k \in A$ until reading depth r, $A^1_{a_k}, \ldots, A^r_{a_k}$.

At the end of each iteration, a division process is executed in the minimal tree. This division takes into account the changes in the densities of the different connections of the neural network that present high average distance. The changes in the densities are determined by means of the relation that exists between two distances: the distance between the neurons and the distance indicated by the minimal tree. Once the clustering has finished, the classification of the new patients can be obtained easily.

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 re-use phases are carried out again so that the system can be ready for the next classification. If 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. 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 re-use stages and extracts knowledge, which is formalized in the set of rules. In this way, human experts 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 [2] 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) [13], although currently CART is the most commonly used.

The results provided by the decision rules are represented graphically. A novel graphical representation has been developed in order to easily identify those individuals not abiding the rules and, as a consequence, to facilitate an easy analysis of the results.

3. Case study: classification of leukaemia patients

The Cancer Institute in the city of Salamanca was interested in novel tools for decision support in the process of leukaemia 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 to incorporate innovative techniques to reduce the dimensionality of the data and to 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 leukaemia: acute lymphocytic leukaemia (ALL), acute myeloid leukaemia (AML), chronic lymphocytic leukaemia (CLL), chronic myeloid leukaemia (CML), and myelodysplastic syndromes (MDS). The aim of the tests performed was to determine whether the system is able to classify new patients based on the cases previously analysed and stored.

Figure 1 represents the model intended to resolve the problem of leukaemia 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 that is external to the model. This phase consists of a set of tests that have been carried out by laboratory personnel and allows 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.

4. Results

The Cancer Institute at Salamanca was interested in establishing new analysis processes in its databases able to incorporate new data mining techniques aimed at reducing the dimensionality of the data, clustering the data, and extracting knowledge. The data available were 212 expression arrays from the chip HG U-133 plus 2.0. These data were stored in a database containing information from patients affected by leukaemia and were initially pre-classified. This was the reference used to measure the classification accuracy.

The approach presented in this paper was applied to the leukaemia data provided by the Cancer Institute. Concretely, the number of initial cases was 212, with a total of 54,675 probes. The number

of probes was reduced to 5479 after executing the filtering techniques, with a low variability for the grouping techniques carried out with techniques such as dendograms or PAM.

Table 1 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 1 are those obtained for the classification provided by the SODTNN obtained from the classification of each of the patients.

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 [12], and the results obtained are shown in Figure 3. At the top of Figure 3 can be seen stacked bars, each of them representing an individual. The bars are divided into as many fragments as probes were retrieved using CART. The amplitude of a rectangle represents the magnitude. At the top of the bars, two rectangles are shown, one is used to represent the real group and the other to represent the group estimated using the CART. At the bottom of Figure 3 can be seen parallel coordinates. Each of the bars represents one of the probes, each of the lines represents the individuals, and the colour indicates the group. The bottom right of Figure 3 presents the probes coloured as shown in the stacked bars. These probes represent the more relevant patterns detected in the cluster by the CART technique. The data have been filtered for the probe in the first coordinate, retrieved using CART. In this way, it is possible to determine the atypical individuals in a simple manner, focusing on the activation or deactivation.

Table 1. Classification provided by the CART rules.

	Total	Successful	Success rate (%)
ALL	11	8	72.73
AML	53	47	88.68
CLL	95	84	88.42
CML	26	22	84.62
MDS	47	43	91.49

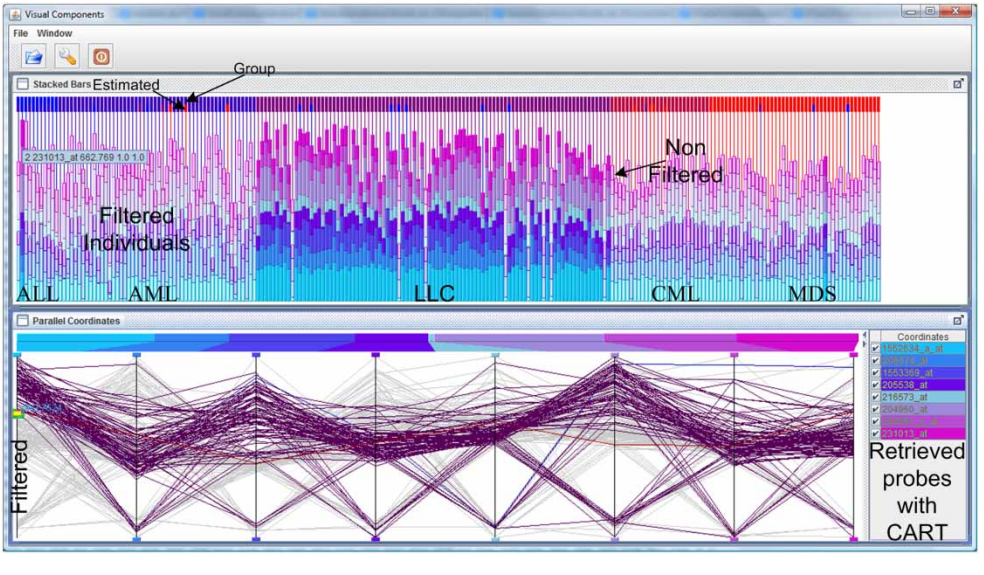

Figure 3. Knowledge extraction with CART.

5. Conclusions

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 on 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. The aim of this study is to optimize the detection of relevant probes. Existing techniques, such as analysis of variance, are based on hypothesis contrast and make use of box plots to represent the results, so that the probes selection process is highly manual and makes necessary to analyse the box plots one by one. The results obtained in this work are promising. However, it is necessary to continue investigating in new organizational techniques to obtain more realistic models to simulate the workflow in the expression analysis. Multi-agent systems seem to be more appropriate to achieve this goal, which is our next challenge.

References

[1] Affymetrix. *GeneChip® Human Genome U133 Arrays* (2010). Available at http://www.affymetrix.com/support/technical/datasheets/hgu133arrays_datasheet.pdf.
[2] L. Breiman, J. Friedman, A. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, Monterrey, CA, 1984.
[3] R. Brunelli, *Histogram analysis for image retrieval*, Patt. Recogn. 34 (2001), pp. 1625–1637.
[4] J.M. Corchado, J. Bajo, Y. De Paz, and J.F. De Paz, *Integrating case planning and RPTW neuronal networks to construct an intelligent environment for health care*, Exp. Syst. Appl. 36 (2009), pp. 5844–5858.
[5] J.F. De Paz, S. Rodríguez, J. Bajo, J.M. Corchado, and V. López, *Self Organized Dynamic Tree Neural Network*, IWANN 09: International Work-Conference on Artificial Neural Networks, Salamanca, 2009, pp. 221–228.
[6] S. Furao, T. Ogura, and O. Hasegawa, *An enhanced self-organizing incremental neural network for online unsupervised learning*, Neural Netw. 20 (2007), pp. 893–903.
[7] 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), pp. 249–264.
[8] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
[9] T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biol. Cybern. 43 (1982), pp. 59–69.
[10] J. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann, 1993.
[11] T. Martinetz, *Competitive Hebbian Learning Rule Forms Perfectly Topology Preserving Maps*, ICANN'93: International Conference on Artificial Neural Networks, Istanbul, Turkey, 1993, pp. 427–434.
[12] T. Martinetz and K. Schulten, *A neural-gas network learns topologies*, Artif. Neural Netw. 1 (1991), pp. 397–402.
[13] J. Quinlan, *Discovering rules by induction from large collections of examples*, in *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, 1979, pp. 168–201.
[14] S.J. Redmond and C. Heneghan, *A method for initialising the K-means clustering algorithm using kd-trees*, Patt. Recogn. Lett. 28 (2007), pp. 965–973.
[15] N. Saitou and M. Nie, *The neighbor-joining method: A new method for reconstructing phylogenetic trees*, *Mol. Biol. Evol*. 4 (1987), pp. 406–425.
[16] N.L.W. van Hal, O. Vorst, A.M.M.L. van Houwelingen, E.J. Kok, A. Peijnenburg, A. Aharoni, A.J. van Tunen, and J. Keijer, *The application of DNA microarrays in gene expression analysis*, J. Biotechnol. 78 (2000), pp. 271–280.
[17] C.K. Yoo, I.B. Lee, and P.A. Vanrolleghem, *Interpreting patterns and analysis of acute leukemia gene expression data by multivariate fuzzy statistical analysis*, Comp. Chem. Eng. 29 (2005), pp. 1345–1356.