CBR System for Diagnosis of Patients

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

Abstract

Microarray technology allows to measure the expression levels of thousands of genes in an experiment. The use of computational methods is fundamental in cancer research. One of the possibilities is the use of Artificial Intelligence techniques. Several of these techniques have been used to analyze expression arrays. This paper presents a Case-based reasoning (CBR) system for automatic classification of leukemia patients from microarray Data. The system incorporates novel algorithms for data mining that allow to filter and classify as well as extraction of knowledge. The system has been tested and the results obtained are presented in this paper.

1. Introduction

Microarray has become an essential tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease [1]. Microarray technology is based on a database of gene fragments called expressed sequence tags (ESTs), which are used to measure target abundance using the scanned intensities of fluorescence from tagged molecules hybridized to ESTs [2]. Specifically, the HG U133 plus 2.0 are chips used for this kind of analysis. These chips analyze the expression level of over 47.000 transcripts and variants, including 38.500 well-characterized human genes. It is comprised of more than 54.000 probe sets and 1.300.000 distinct oligonucleotide feature. The HG U133 plus 2.0 provides multiple, independent measurements for each transcript. Multiple probes mean you get a complete data set with accurate, reliable, reproducible results from every experiment.

The process of studying a microarray is called expression analysis and consists of a series of phases: data collection, data pre-processing, statistical analysis, and biological interpretation. These phases analysis consists basically of three stages: normalization and filtering; clustering and classification. These stages can be automated and included in a CBR [3] system. The first step is critical to achieve both a good normalization of data and an initial filtering to reduce the dimensionality of the data set with which to work [4]. Since the problem at hand is working with high-dimensional arrays, it is important to have a good pre-processing technique that facilitates automatic decision-making about the variables that will be vital for the classification process. In light of these decisions it will be possible to reduce the original dataset. Moreover, the choice of a clustering technique allows data to be grouped according to certain variables that dominate the behaviour of the group. After organizing into groups it is possible to extract of knowledge and classify patients within the group which presents the most similarities.

This technology has been adopted by the research community for the study of a wide range of biologic processes allowing carry out diagnosis. Currently, it is being very used [1] for diagnosing of cancer such as Leukemias. Leukemia, or blood cancer, is a disease that has a significant potential for cure if detected early [5]. The relationship between the chromosomal alterations and prognosis of leukemia and lymphomas is well established. Recently, conventional array-based expression profiling has demonstrated that chromosomal alterations are associated with distinctive patterns of expression. Leukemia is a blood cancer form, originating in a malfunctioning bone marrow that tends to produce abnormal red and white cells at an increased rate. The system proposed in the context of this work focuses on the detection of carcinogenic patterns in the data from microarrays for patients, and is constructed from a CBR system that provides a classification technique based on previous experiences.

For some time now, we have been working on the identification of techniques to automate the reasoning cycle of several CBR systems applied to complex domains [3]. The objective of this work is to develop a

CBR system that allows the identification of patients with various types of cancer. The model aims to improve the cancer classification based on microarray data. The system proposed in this paper presents a new synthesis that brings several artificial intelligence subfields together (filter techniques, clustering, artificial neural networks and extraction of knowledge). The retrieval, reuse, revision and learning stages of the CBR system use these techniques to facilitate the CBR adaptation to the domain of biological discovery with microarray datasets. Specifically, the system presented in this paper uses a model which takes advantage of two novel methods for analyzing microarray data: a technique for filtering data, and a technique ESOINN [25] for clustering. The first method combines various filtering techniques to dramatically reduce the dimensionality of the data. The second one allows clustering by incorporating both the distribution process of the entire surface of classification, and the separation between groups with low density among them.

The paper is structured as follows: The next section presents the problem that motivates this research, i.e., the classification of leukemia patients from samples obtained through microarrays. Section 2 describe the proposed CBR model and how it is adapted to the problem under consideration. Finally, Section 3 presents the results and conclusions obtained after testing the model.

2. CBR System for Classifying Microarray Data

The CBR developed tool receives data from the analysis of chips and is responsible for classifying of individuals based on evidence and existing data. The purpose of case-based reasoning (CBR) is to solve new problems by adapting solutions that have been used to solve similar problems in the past [6]. The primary concept when working with CBRs is the concept of case. A case can be defined as a past experience, and is composed of three elements: A problem description which describes the initial problem, a solution which provides the sequence of actions carried out in order to solve the problem, and the final state which describes the state achieved once the solution was applied. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential steps which are recalled every time that a problem needs to be solved: retrieve, reuse, revise and retain. Each of the steps of the CBR life cycle requires a model or method in order to perform its mission. The algorithms selected for the retrieval of cases should be able to search the case base and to select from it the most similar problems, together with their solutions, to the new problem. In our case study, it conducted a filtering of variables, recovering important variables of the cases to determine the most influential in the conduct classification. Once the most important variables have been retrieved, the reuse phase begins, adapting the solutions for the retrieved cases to obtain the clustering. Once this grouping is accomplished, the next step is to extract the knowledge. The revise phase consists of an expert revision for the solution proposed, and finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases, consequently updating the cases memory.

2.1. Retrieve

Traditionally, only the similar cases to the current problem are recovered, often because of performance, and then adapted. In the case study, the number of cases is not the problem, rather the number of variables. For this reason variables are retrieved at this stage and then, depending on the identified variables, the other stages of the CBR are carried out. This phase will be broken down into 6 stages which are described below:

2.1.1. RMA: The RMA (Robust Multi-array Average) [7] algorithm is frequently used for pre-processing Affymetrix microarray data. RMA consists of three steps: (i) Background Correction; (ii) Quantile Normalization (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.

2.1.2. 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.

2.1.3. 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.

.

2.1.4. Variability: Once both the control and the erroneous probes have been eliminated, the filtering begins. The first stage is to remove the probes that have low variability. This work is carried out according to the following steps:

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

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

   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

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

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

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

2.1.5. Uniform distribution: Finally, all remaining variables that follow a uniform distribution are eliminated. The variables that follow a uniform distribution will not allow the separation of individuals. Therefore, the variables that do not follow this distribution will be really useful variables in the classification of the cases. The contrast of assumptions followed is explained below, using the Kolmogorov-Smirnov [8] test as an example. H0: The data follow a uniform distribution; H1: The analyzed data do not follow a uniform distribution. Statistical contrast:

$$D = \max\left\{D^{+}, D^{-}\right\} \qquad (3)$$

where

$$D^{+} = \max_{1\le i\le n}\left\{\frac{i}{n} - F_0(x_i)\right\}$$

$$D^{-} = \max_{1\le i\le n}\left\{F_0(x_i) - \frac{i-1}{n}\right\}$$

with i as the pattern of entry, n the number of items and $F_0(x_i)$ the probability of observing values less than i with H_0 being true. The value of statistical contrast is compared to the next value:

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

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

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

$$r_{x_{\cdot i}y_{\cdot j}} > \alpha \qquad (5)$$

being: $\alpha = 0.95$ $\quad r_{x_{\cdot i}y_{\cdot j}} = \frac{\sigma_{x_{\cdot i}x_{\cdot j}}}{\sigma_{x_{\cdot i}}\sigma_{x_{\cdot j}}}$,

$\sigma_{x_{\cdot i}x_{\cdot j}} = \frac{1}{N}\sum_{s=1}^{N}\left(\bar{\mu}_{\cdot i} - x_{si}\right)\left(\bar{u}_{\cdot j} - x_{sj}\right)$ Where $\sigma_{x_{\cdot i}x_{\cdot j}}$ is the covariance between probes i and j.

2.2. Reuse

Once filtered and standardized, the probes produce a set of values with i = 1 ... N, j = 1 ... s where N is the total number of cases, s the number of end probes. The next step is to perform the clustering of individuals based on their proximity according to their probes. 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 SOM [9] (self-organizing map), GNG [10] (Growing neural Gas) resulting from the union of techniques CHL [11] (competitive Hebbian Learning) and NG [12] (neural gas), GCS [11] (Growing Cell Structure), Growing Grid or the SOINN [13] (self-organizing incremental neuronal network). 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 behaviour on methods similar to the NG. They create a mesh that is adjusted automatically to a specific area. The greatest disadvantage, however, is that both the number of neurons that are distributed over the surface and the degree of proximity are set beforehand, resulting in the number remaining constant throughout the entire training process, thus complicating, to a certain extend, the adaptation of the mesh. Unlike the self-organizing maps based on meshes, Growing Grid or GCS do not set the number of neurons, or the degree of connectivity, but they do establish the dimensionality of each mesh. This complicates the separation phase between groups once it is distributed evenly across the surface.

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 neural network SOINN [13], called ESOINN [14] (Enhanced self-organizing incremental neuronal network). Unlike the SOINN, ESOINN consists of a single layer, so it is not necessary to determine the manner in which the training of the first layer changes to the second. With a single layer, ESOINN is able to incorporate both the distribution process along the surface and the separation between low density groups. The operation and training of the network presents many similarities with those used in GCS networks as far as distribution over the surface is concerned, but not as far as the dimensionality of the meshes. Nevertheless, it more closely resembles a merger between a CHL and a NG: it has characteristics of a network CHL in the initial phases of the algorithm, by which it could be understood as a phase of competition, while in a second phase, the network of nodes begins to expand just as with a NG network. This process is conducted in an iterative way until it reaches stability. Only the changes in training phase are detailed below:

1. Update the weights of neurons by following a process similar to the SOINN, but introducing a new definition for the learning rate in order to provide greater stability for the model. This learning rate has produced good results in other networks such as SOM [17].

$$\Delta W_{a_1} = n_1(M_{a_1})(\xi - W_{a_1})$$

$$\Delta W_i = n_1(M_{a_i})(\xi - W_{a_i}) \text{ with } i \in N_i \qquad (6)$$

Being $n_1(x) = \frac{1}{\sqrt{x}}$, $n_2(x) = \frac{1}{\sqrt{2+x^2}}$

2. Delete the connections with higher age. The ages are typified and are removed those whose values are in the region of rejection with k>0. The value of α chosen is 0.05.
3. If all input patterns have been passed then a KS-Test [8] is carried out in order to determine if the density distribution for the neurons in each group follows a normal distribution. If so then the learning procedure is finished; otherwise the next pattern is processed. The value of α chosen is 0.05.

Once the clustering has been generated the extraction of knowledge using the CART [15] algorithm is carried out, and finally the new case is classified. The CART algorithm is a non parametric test that allows extracting rules that explain the classification carried out in the previous steps. There are others techniques to generate the decision trees, that is the case of the methods based on ID3 trees [16], although the most used currently is CART. This method allows to generate rules and to extract the most important variables to classify patients with high performance.

2.3. Revise and Retain

The revision is carried out by an expert who determines the correction with the group assigned by the system. If the assignation is considered correct, then the retrieve and reuse phases are carried out again so that the system is ready for the next classification.

3. Results and Conclusions

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

In the study of leukemia on the basis of data from microarrays, 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 212 adult patients with five types of leukaemia. The retrieve stage of the proposed CBR system presents a novel technique to reduce the dimensionality of the data. The total number of probes in our experiments was reduced to 785, which increased the efficiency of the cluster probe. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer. The error rates have remained fairly low especially for cases where the number of patients was high. To try to increase the reduction of the dimensionality of the data we applied principal components (PCA) [18], following the method of Eigen values over 1. A total of 93 factors were generated, collecting 96% of the variability. However, this reduction of the dimensionality was not appropriate in order to obtain a correct classification of the patients, so this step was removed from the recovery phase. Figure 1 shows the classification performed for patients from groups MDS. As can be seen in Figure 1, represented in black, most of the people of the MDS group are together, coinciding with the previous classification given by the experts at the Institute of Cancer. Only a small portion of the individuals departed from the initial classification. Groups that have fewer individuals are those with a higher classification error.

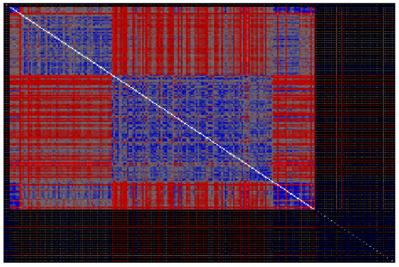


Figure 1. Classification obtained for MDS patients

In a similar way we proceeded to evaluate the classification for the rest of the groups. Table 1 shows the total number of patients from each group and the number of misclassifications. As can be seen in Table 1, groups with fewer patients are those with a greater error rate.

Table 1. Classification errors numerical

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

The final classification was compared with the data obtained using a dendogram [21] and PAM [20] (Partitioning Around Medoids). The proportion of errors in every group was calculated and the Kurskal-Wallis [19] test was applied to determinate if the median of these proportions was equal. The results are shown in table 2.

Table 2. Comparison of methods. * different median and = equal, (-) median of column less than median of row

	CBR	Dendogram	PAM
CBR			
Dendogram	*(-)		
PAM	*(-)	*(-)	

Once checked that the retrieved probes allow classifying the patients in similar way to the original one, we can conclude that the retrieve phase works satisfactorily. Then, the extraction of knowledge is carried out bearing in mind the selected probes. The algorithm used was CART [12], and the results obtained are shown in Figure 2.

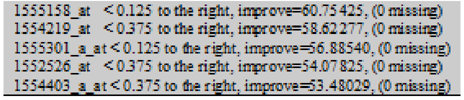
```
n= 212

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

 1) root 212 123 CLL (0.047 0.23 0.42 0.1 0.2)
   2) x1555158_at>=0.125 90   4 CLL (0.033 0 0.96 0.011 0) *
   3) x1555158_at< 0.125 122  73 AML (0.057 0.4 0.025 0.17 0.34)
     6) x1555139_a_at>=0.375 44   7 AML (0.068 0.84 0.023 0.023 0.045) *
     7) x1555139_a_at< 0.375 78  38 DSM (0.051 0.15 0.026 0.26 0.51)
      14) x1554827_a_at< 0.625 30  12 CML (0.067 0.17 0.067 0.6 0.1)
        28) x1554696_s_at>=0.375 8   3 AML (0.25 0.62 0 0 0.12) *
        29) x1554696_s_at< 0.375 22   4 CML (0 0 0.091 0.82 0.091) *
      15) x1554827_a_at>=0.625 48  11 MDS (0.042 0.15 0 0.042 0.77)
        30) x1553244_at< 0.125 9   3 AML (0.22 0.67 0 0 0.11) *
          31) x1553244_at>=0.125 39   3 MDS (0 0.026 0 0.051 0.92) *
```

Figure 2. Extraction of knowledge

The most important probes and their relevance in the classification of patients are extracted by means of this algorithm. In Figure 3 the most important probes in CLL are shown:



```
1555158_at   < 0.125 to the right, improve=60.75425, (0 missing)
1554219_at   < 0.375 to the right, improve=58.62277, (0 missing)
1555301_a_at < 0.125 to the right, improve=56.88540, (0 missing)
1552526_at   < 0.375 to the right, improve=54.07825, (0 missing)
1554403_a_at < 0.375 to the right, improve=53.48029, (0 missing)
```

Figure 3. Extraction of knowledge CLL

The proposed model resolves this problem by using a technique that detects the genes of importance for the classification of diseases by analysing 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 technique for clustering based in neuronal networks. The results obtained from empirical studies provide a tool that allows both the detection of genes and those variables that are most important for the detection of pathology, and the facilitation of a classification and reliable diagnosis, as shown by the results presented in this paper.

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

4. References

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

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

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

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

[5] J.E. Rubnitz, N. Hijiya, Y. Zhou, M.L. Hancock, G.K. Rivera, C. Pui, "Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia." *Pediatric Blood & Cancer*, Vol. 44 (2), (2005) pp. 138-141

[6] J. Kolodner, "Case-Based Reasoning." *Morgan Kaufmann* (1993)

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

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

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

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

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

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

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

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

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

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

[17] J.M. Corchado, J. Bajo, Y. De Paz, J.F. De Paz "Integrating Case Planning and RPTW Neuronal Networks to Construct an Intelligent Environment for Health Care." *Expert Systems with Applications*, In Press (2008)

[18] I. Jolliffe, "Principal Component Analysis." *Springer Series in Statistics* (2002)

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

[20] L. Kaufman, P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis." Wiley, New York. (1990)

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