Applying CBR Systems to Micro Array Data Classification

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

Departamento de Informática y Automática, Universidad de Salamanca Plaza de la Merced s/n, 37008, Salamanca, Spain
{srg,fcofds,jbajope,corchado}@usal.es

Summary. Microarray technology allows to measureing the expression levels of thousands of genes in an experiment. This technology required requires computational solutions capable of dealing with great amounts of data and as well as techniques to explore the data and extract knowledge which allow patients classification. This paper presents a systems based on Case-based reasoning (CBR) 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.

Key words: Case-based Reasoning, HG U133, dendogram, leukemia classification, decision tree

1 Introduction

The progress in the biomedicine [1] [24] and the incorporation of computational and artificial intelligence techniques, have caused a big progress in the detection and diagnosis of many illness. Microarray technology allows to measure the expression levels of thousands of genes in an experiment. 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 [23] for diagnosing of cancer such as Leukemias. This technique studies RNA chains thereby identifying the level of expression for each gene studied. It consists of hybridizing a sample for a patient and colouring the cellular material with a special dye. This offers different levels of luminescence that can be analyzed and represented as a data array [24]. These methods and tools need to work with expression arrays containing a large amount of data points. 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. Eleven pairs of oligonucleotide probes are used to measure the level of transcription of each sequence represented on the GeneChip Human Genome Focus Array.

The process of studying a microarray is called expression analysis and consists of a series of phases: data collection, data preprocessing, statistical analysis, and biological interpretation. These phases analysis consists basically of three stages: normalization and filtering; clustering and classification; and extraction of knowledge. These stages can be automated and included in a CBR [15] 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 [5]. Moreover, the choice of a clustering technique allows data to be grouped according to certain variables that dominate the behaviour of the group [6]. After organizing into groups it is possible to extract the information necessary about the most significant probes that characterize each cluster. In base on this information, the association of new individuals to a cluster can be carried out. Finally, experts can learn from the analysis process.

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 [22] [15]. The microarray analysis to distinguish subclasses in disease and identify pattern associated with disease according to its genes. This patterns of expression that are used to classify types leukimia. This paper presents a CBR system that facilitates the analysis and classification of data from microarrays corresponding to patients with leukemia. Leukemia, or blood cancer, is a disease that has a significant potential for cure if detected early [4]. The model aims to improve the cancer classification based on microarray data using CBR. The system presented in this paper uses a model which takes advantage of three methods for analyzing microarray data: a technique for filtering data, a technique for clustering and a method for extracting the knowledge.

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 and Section 3 describe the proposed CBR model and how it is adapted to the problem under consideration. Finally, Section 4 presents the results and conclusions obtained after testing the model.

2 CBR System for Classifying Micro Array 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. Case-based Reasoning is a type of reasoning based on the use of past experiences [7]. CBR systems solve new problems by adapting solutions that have

been used to solve similar problems in the past, and learning from each new experience. 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 delineates the initial problem; a solution, which provides the sequence of actions carried out in order to solve the problem; and the final stage, which describes the state achieved once the solution was applied. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain. The retrieve phase starts when a new problem description is received. In this phase a similarity algorithm is used to find the greatest number of cases in the cases memory. 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 as explained in section 2.1. 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 determine the provenance of the new individual to be evaluated. 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

Contrary to what usually happens in the CBR, our case study is unique in that the number of variables is much greater than the number of cases. This leads to a change in the way the CBR functions so that instead of recovering cases at this stage, important variables are retrieved. 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 5 stages which are described below:

RMA: The RMA (Robust Multi-array Average) [8] 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. 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

$$x_{i,j,n} = \mu_{i,n} + \alpha_{j,n} + \epsilon_{i,j,n} \; with \; i = 1...I, j = 1...J, n = 1...N \sum_j \alpha_j = 0 \quad (1)$$

Where α_j is a probe affinity effect, μ_j represents the log_2 scale expression level for array i and $\epsilon_{i,j}$ represents an independent identically distributed error term with mean 0. Median polish [21] is used to obtain estimates of the values.

Control and error: 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. 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.

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

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

   Where N is the number of items total, is the average population for the variable j, is the value of the probe j for the individual i.
2. Standardize the above values

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

   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: $z < -1.0$ given that $P(z < -1.0) = 0.1587$. This will effect the removal of about 16% of the probes if the variable follows a normal distribution.

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 [13] test as an example.

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

where $D^+ = max_{1\leq i\leq n}\left\{\frac{i}{n} - F_o\left(x_i\right)\right\}$ $D^- = max_{1\leq i\leq n}\left\{F_o\left(x_i\right) - \frac{i-1}{n}\right\}$ with i as the pattern of entry, n the number of items and $F_o\left(x_i\right)$ the probability

of observing values less than i with H_o 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 $k\,(n) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$ and a level of significance $\alpha = 0.05$ $C_\alpha = 1.358$.

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.

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

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

2.2 Reuse

Once filtered and standardized the data using different techniques of data mining, the system produce a set of values x_{ij} 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 in data mining. 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 [10] (Growing Cell Structure). There are other techniques with less computational cost that provide efficient results. Among them we can find the dendogram and the PAM method [16] (Partitioning Around Medoids). A dendrogram [17] is a ascendant hierarchical method with graphical representation that facilitates the interpretation of results and allows an easy way to establish groups without prior establishment. The PAM method requires a selection of the number of clusters previous to its execution.

The dendograms are hierarchical methods that initially define as conglomerates for each available cases. At each stage the method joins those conglomerates of smaller distance and calculates the distance of the conglomerate with everyone else. The new distances are updated in the matrix of distances. The process finishes when there is one only conglomerate (agglomerative method). The distance metric used in this paper has been the average linkage. This metric calculates the average distance of each pair of nodes for the two groups,

and based on these distances mergers the groups. The metric is known as unweighted pair group method using arithmetic averages (UPGMA) [18]. Once the dendogram has been generated, the error rate is calculated bearing in mind the previous cases. If the accuracy rate is up to 80%, the extraction of knowledge using the CART (Classification and Regression Tree) [19] 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 [20], 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 Case Study

In the case study presented in the framework of this research are available 232 samples are available from analyses performed on patients either through punctures in marrow or blood samples. The aim of the tests performed is to determine whether the system is able to classify new patients based on the previous cases analyzed and stored.

Figure 1 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. As can be seen in Figure 1, a previous phase, external to the model, 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 luminescence. At that point, the CBR-based model starts to process the data obtained from the microarrays.

The retrieve 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 retrieve 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 $D = \{d_1, ..., d_t\}$, where $d_i \in \mathbb{R}^n$ represents the patient i and n represents the number of probes taken into consideration. As explained in Section 2.1 during the retrieve phase the data are normalized by the RMA algorithm [8] and the dimensionality is reduced bearing

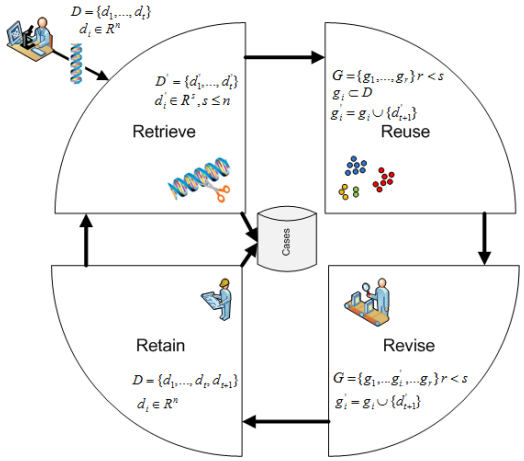

Fig. 1. Proposed CBR model

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 $D^{'} = \left\{d_1^{'}, ..., d_t^{'}\right\}\ d_i^{'} \in \mathbb{R}^s, s \leq n$.

The reuse phase 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 retriever phase consists of a group of patients $D^{'} = \left\{d_1^{'}, ..., d_t^{'}\right\}$ with $d_i^{'} \in \mathbb{R}^s, s \leq n$ each one characterized by a set of meaningful attributes $d_i = (x_{i1}, ..., x : is)$, where x_{ij} is the luminescence value of the probe i for the patient j. In order to create clusters and consequently obtain patterns to classify the new patient, the reuse phase implements a method of hierarchical cluster called dendogram, which has been explained in section 2.2. The system 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 $G = \{g_1, ..., g_r\}\ r < s\ g_i \subset D,\ g_i \cap g_j = \phi$ 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 accuracy rate is calculated, if it is greater than 80% then the clustering and extraction of knowledge are carried out. The new patient is defined as $d_{t+1}^{'}$ and his membership to a group is determined following the classification tree in section 2.2. The result of the reuse phase is a group of clusters $G = \left\{g_1, ..., g_i^{'}, ..., g_r\right\}\ r < s$ where $g_i^{'} = g_i \cup \left\{d_{t+1}^{'}\right\}$.

An expert from the Cancer Institute is in charge of the revision process. This expert determines if $g_i^{'} = g_i \cup \left\{d_{t+1}^{'}\right\}$ can be considered as correct. In the retain phase the system learns from the new experience. If the classification

is considered successful, then the patient is added to the memory case $D = \{d_1, ..., d_t, d_{t+1}\}$.

4 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 method of clustering and extraction the knowledge. 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 variables selected 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) [14], 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. Figure 2a shows the classification performed for patients from all the groups. In the left it is possible to observe the groups identified in the classification process. Cases interspersed represent individuals with different classification to the previous-one. As shown in Figure 2a the number of misclassified individuals have been low.

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 [19] and the results obtained are shown in Figure 2b.

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

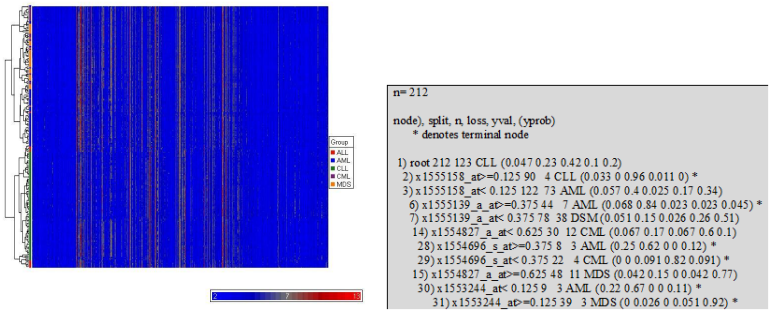

Fig. 2. Classification obtained.

data. It also presents a technique for clustering based in hierarchical methods. The results obtained from empirical studies are promising and highly appreciated by specialists from the laboratory, as they are provided with 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 for the information and technology provided.

References

1. Shortliffe E, Cimino J (2006) Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer Berlin Heidelberg New York
2. Tsoka S, Ouzounis C (2000) Recent developments and future directions in computational genomics. FEBS Letters 480 (1):42–48
3. Lander E et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
4. Rubnitz J, Hijiya N, Zhou Y, Hancock M, Rivera G, Pui C (2005) Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatric Blood & Cancer 44 (2):138–141
5. Armstrong N, van de Wiel M (2004) Microarray data analysis: From hypotheses to conclusions using gene expression data. Cellular Oncology. 26 (5-6):279-290
6. Quackenbush J (2001) Computational analysis of microarray data. Nature Review Genetics 2(6):418–427
7. Kolodner J (1993) Case-Based Reasoning. Morgan Kaufmann
8. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T (2003) Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. Biostatistics 4:249–264

9. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 59-69
10. Fritzke B (1995) A growing neural gas network learns topologies. Cambridge MA: Tesauro G, Touretzky D, Leen T (eds) Advances in Neural Information Processing Systems 7 625–632
11. Martinetz T (1993) Competitive Hebbian learning rule forms perfectly topology preserving maps. ICANN'93: International Conference on Artificial Neural Networks. Springer Amsterdam 427–434
12. Martinetz T, Schulten K (1991) A neural-gas network learns topologies (1991) Kohonen T, Makisara K, Simula 0, Kangas J (eds) Artificial Neural Networks Amsterdam 397-402
13. Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition, Vol. 34, (2001) 1625–1637
14. Jolliffe I (2002) Principal Component Analysis. Second Edition. Springer Series in Statistics
15. Riverola F, Daz F, Corchado J (2006) Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence 22(3-4):254–268
16. Kaufman L, Rousseeuw P (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York
17. Saitou N, Nie M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425
18. Sneath P, Sokal R (1973) Numerical Taxonomy. The Principles and Practice of Numerical Classification. W.H. Freeman Company, San Francisco
19. Breiman L, Friedman J, Olshen A, Stone C (1984) Classification and regression trees. Wadsworth International Group. Belmont, California
20. Quinlan J (1979) Discovering rules by induction from large collections of examples. In D. Michie Eds., Expert systems in the micro electronic age. Edinburgh University Press. Edinburgh. 168–201
21. Holder D, Raubertas R, Pikounis V, Svetnik V and Soper K (2001) Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In Proceedings of the ASA Annual Meeting Atlanta, GA
22. Corchado J, Corchado E, Aiken J, Fyfe C, Fdez-Riverola F, Glez-Bedia M (2003) Maximum Likelihood Hebbian Learning Based Retrieval Method for CBR Systems. In Proceedings. of the 5th International Conference on Case-Based Reasoning. 107–121
23. Quackenbush J (2006) Microarray Analysis and Tumor Classification, The new england journal o f medicine. 2463–2472
24. Zhenyu C, Jianping L, Liwei W, (2007) A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artificial Intelligence in Medicine 41:161–175