CBR System with Reinforce in the Revision Phase for the Classification of CLL Leukemia

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 measuring the expression levels of thousands of genes providing huge quantities of data to be analyzed. This fact makes fundamental the use of computational methods as well as new intelligent algorithms. This paper presents a Case-based reasoning (CBR) system for automatic classification of microarray data. The CBR system incorporates novel algorithms for data classification and knowledge discovery. The system has been tested in a case study and the results obtained are presented.

Keywords: Case-based Reasoning, CLL, luekemia, HG U133.

1 Introduction

The use of microarrays, and more specifically expression arrays, enables the analysis of different sequences of oligonucleotides [1], [2]. Simply put a microarray is an array of probes that contains genetic material with a predetermined sequence. These sequences are hybridized with the genetic material of patients, thus allowing the detection of genetic mutations through the analysis of the presence or absence of certain sequences of genetic material. This work focuses on the levels of expression for the different genes, as well as on the identification of the probes that characterize the genes and allow the classification into groups.

The analysis of expression arrays is called expression analysis. An expression analysis basically consists of three stages: normalization and filtering; clustering and classification; and extraction of knowledge. These stages are carried out from the luminescence values found in the probes. Presently, the number of probes containing expression arrays has increased considerably to the extent that it has become necessary to use new methods and techniques to analyze the information more efficiently. There are various artificial intelligence techniques such as artificial neural networks [4], [5], Bayesian networks [6], and fuzzy logic [7] which have been applied to microarray analysis. While these techniques can be applied at various stages of expression analysis, the knowledge obtained cannot be incorporated into successive tests and included in subsequent analyses.

This paper presents a system based on CBR which uses past experiences to solve new problems [8], [9]. As such, it is perfectly suited for solving the problem at hand. In addition, CBR makes it possible to incorporate the various stages of expression analysis into the reasoning cycle of the CBR, thus facilitating the creation of

S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, pp. 963–970, 2009.

strategies similar to the processes followed in medical laboratories. The recovery of information from previous analyses simplifies the classification process by detecting and eliminating relevant and irrelevant probes detected in previous analyses. This system is applied to the classification of subtypes of leukemia, specifically, to detect patterns and extract subgroups within the CLL type of leukemia obtained from the HG U133 plus [3] expression arrays. The CBR system is based on previous works [10], [11] but the stages of the CBR cycle have been modified. The system incorporates new techniques for improving the filtering in the retrieval stage; The reuse phase now includes a parallel execution of the ESOINN (Enhanced Self-Organizing Incremental Neural Network) [12] neural network and the PAM [13], which facilitates a better evaluation of the classification provided; the revise phase includes a MDS (Multidimensional Scaling) technique [14] to obtain representations at low dimensionality.

The paper is structured as follows: Section 2 presents 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 study. Section 4 presents the results and conclusions obtained after testing the model.

2 Microarray Data Analyses

This section presents the CBR system proposed in the context of this research and provides a classification technique based on previous experiences for data from microarrays. The system receives data from the analysis of chips and is responsible for classifying individuals based on evidence and existing data. 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 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 steps which are recalled every time a problem needs to be solved: retrieve, reuse, revise and retain.

The CBR system previously designed [10] has been modified and incorporate new techniques in each of the stages of the CBR cycle. A new rules memory was incorporated to the CBR to store the rules obtained during the revise stage. The retrieve phase, includes a new additional step. Once the most important variables have been retrieved, the reuse phase begins adapting the solutions for the retrieved cases to obtain the clustering. The clustering is now obtained by means of two different techniques, which results are compared. Once this grouping is accomplished, the next step is knowledge extraction. The revise phase consists of an expert revision for the proposed solution. To facilitate this task, we have incorporated a MDS, and finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases, consequently updating the cases memory and the new rules memory.

2.1 Retrieve

Traditionally, only the cases similar to the current problem are recovered, often because of performance, and then adapted. With regards to expression array, the number of cases is not a critical factor, rather the number of variables. For this reason, we have incorporated an innovative strategy where variables are retrieved at this stage and then, depending on the identified variables, the rest of the stages of the CBR are carried out. In the previous version of the CBR, the retriever phase consisted of: RMA, remove irrelevant probes, low variability, uniform distribution and correlations. A new step has been incorporated in the filtering to obtain a more effective reduction of the number of the significant probes.

2.1.1 Cut-Off Points

This step removes the probes that, despite not following a uniform distribution, have no separation between elements, and do not allow the elements to be partitioned. The way to remove the probes is to detect changes in the densities of the data, and to select the final probes. The probes in which cut-offs or high densities are not detected are eliminated, as they do not provide useful information to the classification process. This will keep the probes that allow the separation of individuals. The detection of the separation intervals is performed by calculating the distance between adjacent individuals. Once the distance is calculated, it is possible to determine the potentially relevant values. The selection is carried out by applying confidence intervals for the values of these differences if the values follow a uniform distribution, or by selecting the values above a certain percentile if the values do not follow a normal distribution. This process is formalized as follows:

1. Let I' be the set of individuals with filtered probes together with the new individual, where $x_{\cdot j}$ represents the probe j for all the individuals, and x_{ij} the individual i for the probe j
2. Select the probe $j=1$, $x_{\cdot j}$
3. Sort in increasing order values $x_{\cdot j}$
4. Calculate the value for $x'_{ij} = x_{i+1j} - x_{ij}$
5. Determine if the variable x'_{ij} follows a uniform distribution by means of the Shapiro-Wilk test [19], otherwise go to step 10.
6. Calculate the value for $\overline{x}'_{\cdot j}$
7. Establish the confidence interval for the variance, which is established as $\sigma'^{2}_{\cdot j} \in \left[\frac{(n-1)\cdot S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)\cdot S^2}{\chi^2_{n-1,\alpha/2}} \right]$ with $\alpha = 0.05$ and $n = \# x'_{\cdot j}$ and the number of elements for $x'_{\cdot j}$, S is the sampling variance.
8. Establish the set of elements form x'_{ij} not belonging to the set $Q_j = \{x'_{ij} / x'_{ij} \notin I_{\sigma'_{\cdot j}}\}$
9. Go to step 11.

10. Select those values up to the percentile P_α from every $x'_{\cdot j}$ and establish the set $Q_j = \left\{x'_{ij} / x'_{ij} > P_\alpha\right\}$
11. Select the probe $j+1$ in the case of more probes needing revision and go to step 2.
12. Create the new set of probes $I' = \bigcup x'_{\cdot j} / \exists x'_{ij} \in Q_j / i > \# x' \cdot u \wedge i < \# x' - \# x' \cdot u$
13. Finalize and return the new set of individuals with the filtered probes I'

2.2 Reuse

The reuse phase has been modified: The classification was performed by an ESOINN neural network. The new reuse stage incorporates a PAM technique. Both PAM and ESOINN are executed simultaneously, and their results are compared.

2.2.1 PAM

The PAM algorithm [13] is executed parallel to the clustering in order to facilitate a comparison of the results obtained. The classification made by both methods, PAM and ESOINN, generates an equivalence index between the two methods that determines the consistency of the reuse phase. The algorithm used for PAM is as follows:

1. Select the number of clusters depending on $\#C$, where C is the set of different classes in I'
2. Classify the patients taking all of the variables into account, without any filtering $G^P = \{g^P / g^P \subseteq I\}$ with $g_i^P \cap g_j^P = \phi, \forall i \neq j$
3. Once the groups G^P are created, an assignation is made following the procedure

$$\max_{c_j \in C} \frac{\# I_{c_j} / g_i^P}{\# I_{c_j}} \cdot \frac{\# I'_{c_j}}{\# I'} \tag{1}$$

where $I'_{c_j} = \{s \in I / s \in c_j\}$, I is the set of initial individuals, c_j is the class j, and I'_{c_j} / g_i^P is the set of individuals from c_j restricted to g_i^P.

2.2.2 Equivalence Index

Once the individuals have been classified using both the PAM and the ESOINN neural networks, the equivalence index for both methods eq is calculated, and the error rate for the ESOINN network is determined as a function of the pre-classified cases. The equivalence index is defined as indicated in (10):

$$eq = \frac{\#\left\{i \in I' / i \in c_j^E \wedge i \in c_j^P\right\}}{\# I'} \tag{2}$$

Where c_j^E, c_j^P representan las clases j bajo la red ESOINN y PAM.

2.2.3 Classification

Once the meshes are generated by the clustering process, previously unclassified individuals are now classified by selecting the nearest mesh. When the mesh has been selected, the case is assigned to the class of the mesh selected.

2.3 Revise

To facilitate the human expert task, the equivalence index and the error rate were calculated in the reuse stage. It is important for the medical human expert to understand the classification process performed in the two previous stages. In this sense, the system provides a knowledge extraction method in the Revise phase. This method analyses the steps followed in the retrieve and reuse stages, and extracts knowledge which is then formalized in rules. In this way, the human expert can easily evaluate the classification and extract conclusions concerning the efficiency of the classification process. It has been incorporated knowledge about the probes that cause erroneous classifications. The knowledge extraction phase detects anomalous classifications, since it accounts for the existence of probes with irrelevant information, or those that were decisive for the misclassification. If the human expert notes that the probes contain irrelevant information, they are marked as irrelevant and not taken into account in the next iteration of the CBR cycle. At this stage, in addition to the representation of the decision tree, a 3D representation with the information retrieved is displayed. The dimensionality is reduced by using MDS [14].

2.4 Retain

If the human expert identifies relevant information at the revise stage, the knowledge is acquired and the information obtained is stored. The information that is stored corresponds to the classifications considered correct, the decision rules generated that are considered relevant, and the probes marked as irrelevant. The information stored is divided into the cases memory and the rules memory.

3 Case Study

Microarray analysis has facilitated the identification of certain characteristic genes in the different variants of leukemia [15]. Cancer experts remark on the importance of the identification of the genes associated to each type of cancer in order to establish the most efficient treatments for the patients [20].

The Institute Cancer Institute of Salamanca provided us with 91 samples of patient data and asked for a tool to provide decision support in the expression array analysis process and to incorporate innovative techniques to reduce the dimensionality of the data and identify the variables with a higher influence in the patient's classification. The samples corresponded to patients affected by chronic lymphocytic leukemia. CLL is a disease of lymphocytes that appear to be mature but are biologically immature. These B lymphocytes arise from a subset of CD5-B cells that appear to have a role in autoimmunity. The pathogenesis of chronic lymphocytic leukemia is likely a multistep process, initially involving a polyclonal expansion of CD5-B cells followed by the transformation of a single cell [17]. CLL is one of four main types of leukemia. About 15.110 new cases of CLL will be diagnosed in 2008. Approximately 90.179 people are currently living with CLL, more than the number of people living with any other type of leukemia. Most people with CLL are at least 50 years old [18]. CLL starts with a change to a single cell called a lymphocyte. Over time, the CLL cells multiply and replace normal lymphocytes in the marrow and lymph nodes. The high number of CLL cells in the marrow may crowd out normal blood-forming cells, and CLL cells are not able to fight off infection like normal lymphocytes do [18]. The aim of the tests performed in this study is to determine whether our system is able to classify new patients based on previously analyzed and stored cases.

4 Results and Conclusions

This paper has presented case-based reasoning system specifically designed to analyze data from microarrays, facilitating the grouping and classification of individuals. Moreover, the system provides an innovative method for exploring the classification process and extracting knowledge in the form of rules which help the human experts to understand the classification process and obtain conclusions about the relevance of the probes. The system has been applied to a real case scenario at the Cancer Institute in the city of Salamanca, for classifying CLL leukemia patients. The human experts in the laboratory have remarked on the advantages of using MicroCBR as a decision support system for CLL classification, and have especially noted the facility in acquiring knowledge and explanations.

Section 3 presented the case study considered in this report, which classified 91 CLL leukemia patients into groups. The aim of the case study was to identify the probes that allow classifying the CLL leukemia patients into subgroups. The pre-processing phase began with 54.675 probes. After the pre-processing phase, the filtering process was applied, notably reducing the probes to 618, without increasing the error rate This stage takes the irrelevant probes filtering into account, as probe 201909_at, which is associated to the Y chromosome.

The reuse phase begins when the probes have been filtered, and generates the meshes for the groups as well as the distribution of the individuals along the space. The mesh closest to the new case was then selected and the classification was made. To evaluate the proposed model, the system classified 90 individuals together with a new individual, and the results obtained were compared to the previous existing classifications. Table 1 shows the number of elements with an erroneous classification. In the revise phase, the CART algorithm is used to obtain both the significant and irrelevant probes and rules. Table 2 shows the final rules obtained.

Table 1. Classification Errors

	Type 1	Type 2	Type 3
Total	46	12	33
Predicted/Erroneous	47/5	9/0	35/4

Table 2. Classification Errors

Class	Rule
C3	219703_at>0.25
	219703_at≤0.25 && 207502_at≤0.5 && 1552619_a_at>0.25
	219703_at≤0.25 && 207502_at≤0.5 && 1552619_a_at≤0.25 && 242601_at>0.25 && 1553586_at≤0
C2	219703_at≤0.25 && 207502_at>0.5
C1	219703_at≤0.25 && 207502_at≤0.5 && 1552619_a_at≤0.25 && 242601_at≤0.25
	219703_at≤0.25 && 207502_at≤0.5 && 1552619_a_at≤0.25 && 242601_at>0.25 && 1553586_at>0

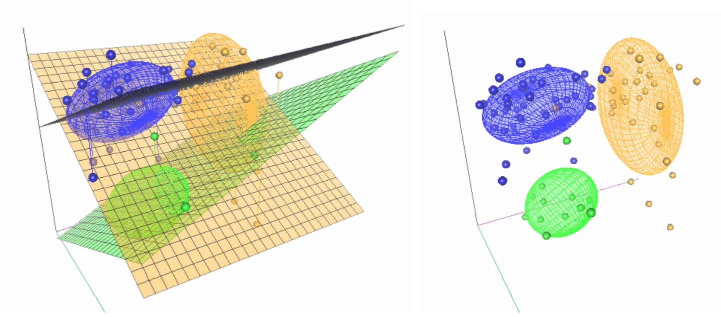

Fig. 1. Representation of low dimensionality probes with MDS

To obtain a visual representation of the patient's classification, we use the MDS [14] and the dimensionality of the data is reduced to three. Figures 1a and 1b represent the information once MDS has been applied and, as shown, the individuals of the different clusters are separated in the space.

The system is able to incorporate the knowledge acquired in previous classifications and use it to perform new classifications, providing a much appreciated decision support tool. As demonstrated, the proposed system reduces 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 neuronal network ESOINN, which is validated with a PAM technique. Finally, the system incorporates a technique for knowledge extraction and presents it to the human experts in a very intuitive format.

Acknowledgments. This development has been supported by the projects SA071A08 and SIAAD-TSI-020100-2008-307.

References

[1] Lina, K.S., Chien, C.F.: Cluster analysis of genome-wide expression data for feature extraction. Expert Systems with Applications 36(2-2), 3327–3335 (2009)

[2] Stadlera, Z.K., Come, S.E.: Review of gene-expression profiling and its clinical use in breast cancer. Critical Reviews in Oncology/Hematology 69(1), 1–11 (2009)

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

[4] Sawa, T., Ohno-Machado, L.: A neural network based similarity index for clustering DNA microarray data. Computers in Biology and Medicine 33(1), 1–15 (2003)

[5] Bianchia, D., Calogero, R., Tirozzi, B.: Kohonen neural networks and genetic classification. Mathematical and Computer Modelling 45(1-2), 34–60 (2007)

[6] Baladandayuthapani, V., Ray, S., Mallick, B.K.: Bayesian Methods for DNA Microarray Data Analysis. Handbook of Statistics 25(1), 713–742 (2005)

[7] Avogadri, R., Valentini, G.: Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artificial Intelligence in Medicine (in press)

[8] Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993)

[9] Riverola, F., Díaz, F., Corchado, J.M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence 22(3-4), 254–268 (2006)

[10] Rodríguez, S., De Paz, J.F., Bajo, J., Corchado, J.M.: Applying CBR Systems to Microarray Data Classification. In: IWPACBB 2008. Advances in Soft Computing, vol. 49, pp. 102–111 (2008)

[11] Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of Experts for Decision Support in the Diagnosis of Leukemia Patients. Artificial Intelligence in Medicine (in press)

[12] Furao, S., Ogura, T., Hasegawa, O.: An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks 20(8), 893–903 (2007)

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

[14] Borg, I., Groenen, P.: Modern multidimensional scaling theory and applications. Springer, New York (1997)

[15] Avogadri, R., Valentini, G.: The Corresponding Author and Giorgio Valentini Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artificial Intelligence in Medicine (in press)

[16] Vogiatzis, D., Tsapatsoulis, N.: Active learning for microarray data. International Journal of Approximate Reasoning 47(1), 85–96 (2008)

[17] Foon, K.A., Rai, K.L., Gale, R.P.: Chronic lymphocytic leukemia: new insights into biology and therapy. Annals of Internal Medicine 113(7), 525–539 (1990)

[18] Chronic Lymphocytic Leukemia. The leukemia and lymphoma society (2008), http://www.leukemia-lymphoma.org/all_page.adp?item_id=7059

[19] Jurečkováa, J., Picek, J.: Shapiro–Wilk type test of normality under nuisance regression and scale. Computational Statistics & Data Analysis 51(10), 5184–5191 (2007)

[20] Yang, T.Y.: Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern. Computational Statistics & Data Analysis (in press)