Using CBR Systems for Leukemia Classification

Juan M. Corchado and Juan F. De Paz

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

Abstract. The continuous advances in genomics, and specifically in the field of transcriptome, require novel computational solutions capable of dealing with great amounts of data. Each expression analysis needs different techniques to explore the data and extract knowledge which allow patients classification. This paper presents a hybrid systems based on Case-based reasoning (CBR) for automatic classification of leukemia patients from Exon array data. The system incorporates novel algorithms for data mining that allow to filter and classify. The system has been tested and the results obtained are presented in this paper.

Keywords: Case-based Reasoning, dendogram, leukemia classification, data mining

1. Introduction

During recent years there have been great advances in the field of Biomedicine [1]. The incorporation of computational and artificial intelligence techniques to the field of medicine has yielded remarkable progress in predicting and detecting diseases [1]. One of the areas of medicine which is essential and requires the implementation of techniques that facilitate automatic data processing is genomics. Genomics deals with the study of genes, their documentation, their structure and how they interact [2]. We distinguish different fields of study within the genome. One is transcriptome, which deals with the study of ribonucleic acid (RNA), and can be studied through expression analysis [3]. 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 represented as a data array. Traditionally, methods and tools have been developed to work with expression arrays containing about 50.000 data points. The emergence of the Exon arrays [5], holds important potential for biomedicine. However, the Exon arrays require novel tools and methods to work with very large (5.500.000) amounts of data.

Microarray technology has been performed for the identification of acute leukemia prognosis. Microarray has become an essential tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease [6]. 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 [6]. 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 [17] system.

This paper presents a hybrid system system that facilitates the analysis and classification of data from Exon arrays corresponding to patients with leukemia. The system is based on a CBR that incorporates techniques of data mining in the different phases. Leukemia, or blood cancer, is a disease that has a significant potential for cure if detected early [4]. The system proposed in the context of this work focuses on the detection of carcinogenic patterns in the data from Exon arrays.

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 Exon arrays. Section 2 and Section 3 describe the proposed hybrid system 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 Exon 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 [8]. 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 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, 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) [9] 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, follow a linear additive mode.

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... 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
2. Standardize the above values
3. Discard of probes for which the values meet the following condition $z < -1.0$.

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 [14] test as an example. The selected level of significance $\alpha = 0.05$.

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. $\alpha = 0.05$.

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 [10] (self-organizing map), GNG [11] (Growing neural Gas) resulting from the union of techniques CHL [12] (Competitive Hebbian Learning) and NG [13] (neural gas), GCS [11] (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 [18] (Partitioning Around Medoids). A dendrogram [19] is a ascendant hierarchical method with graphical representation that facilitates the interpretation of results and allows an easy way to establish groups without prior knowledge. 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) [20]. The clustering obtained is compared to the individuals that have already been classified by an expert. The percentage of error represents the confidence level.

Once, the clusters have been made, the new sample is classified. The KNN algorithm [21] (K-Nearest Neighbour) is used. KNN allows setting probabilistic values based on its neighbours. It is necessary to establish a measure of similarity to calculate the distance between individuals. The similarity measure used is as follows:

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

Where s is the total number variables, n and m the cases, w_i the value obtained in the uniform test and f the Minkowski [15] Distance that is given for the following equation.

$$f(x,y) = \sqrt[p]{\sum_i \left|x_i - y_j\right|^p} \text{ con } x_i, y_j \in R^p \qquad (2)$$

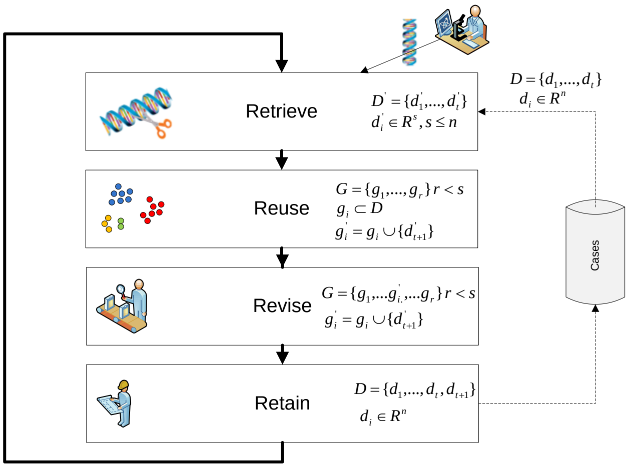
2.3. Revise and Retain

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. If the prognosis of the expert differs from the system, the case is not incorporated until the final medical diagnosis is carried out.

3. Case Study

In the case study presented in the paper, 232 samples were available from analyses performed on patients through punctures in marrow or blood samples, which have been hybridized and analyzed through Exon arrays manufactured by Affymetrix. The aim of the tests performed is to determine whether the system is able to classify new patients based on the previous cases that were 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 Exon arrays.



Retrieve
$D^{'} = \{d_1^{'},...,d_t^{'}\}$
$d_i^{'} \in R^s, s \leq n$
Reuse
$G = \{g_1,...,g_r\} r < s$
$g_i \subset D$
$g_i^{'} = g_i \cup \{d_{t+1}^{'}\}$
Revise
$G = \{g_1,...g_{i.}^{'},...g_r\} r < s$
$g_i^{'} = g_i \cup \{d_{t+1}^{'}\}$
Retain
$D = \{d_1,...,d_t,d_{t+1}\}$
$d_i \in R^n$
$D = \{d_1,...,d_t\}$
$d_i \in R^n$
Cases

Fig. 1. Proposed CBR model

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 $d_i \in R^n$ represents the patient *i* and *n* represents the number of probes taken into consideration and t the number of cases. As explained in

Section 2.1, during the retrieve phase the data are normalized by the RMA algorithm [9] and the dimensionality is reduced bearing in mind, above all, the variability, distribution and correlation of probes. The result of this phase reduces any information not considered meaningful to perform the classification. The new set of patients is defined through *s* variables $D^{'} = \{d_1^{'},...,d_t^{'}\}$ $d_i^{'} \in R^s, s \le n$.

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 $D^{'} = \{d_1^{'},...,d_t^{'}\}$ with $d_i^{'} \in R^s, s \le n$, 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 hierarchical classification method known as dendogram and 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 $i \ne j$ and $i, j < r$. The set G is composed of a group of clusters, each of them containing patients with a similar disease. The clusters have been constructed by taking into account the similarity between the patient's meaningful symptoms. Once the clusters have been obtained, the system can classify the new patient by assigning him to one of the clusters. The new patient is defined as $d_{t+1}^{'}$ and his membership to a group is determined by a similarity function defined in (1). The result of the reuse phase is a group of clusters $G = \{g_1,...g_{i.}^{'},...g_r\} r < s$ where $g_i^{'} = g_i \cup \{d_{t+1}^{'}\}$.

An expert from the Cancer Institute is in charge of the revision process. This expert determines if $g_i^{'} = g_i \cup \{d_{t+1}^{'}\}$ 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 Exon arrays. 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 Exon arrays, the process of filtering data acquires special importance. In the experiments reported in this paper, we worked with a database of bone marrow cases from 232 adult patients with five types of leukaemia. The data consisted of 5.500.000 scanned intensities. The retrieve stage of the proposed CBR system presents a technique to reduce the dimensionality of the data. The total number of variables selected in our experiments was reduced to 883, which increased the efficiency of the cluster probe. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer. To try to increase the reduction of the dimensionality of the data we applied principal components (PCA) [16], following the method of Eigen values over 1. A total of 112 factors were generated, collecting 96% of the variability. However, this reduction of the dimensionality was not appropriate in order to obtain a correct classification of the patients. Figure 2 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 2 the number of misclassified individuals have been low.

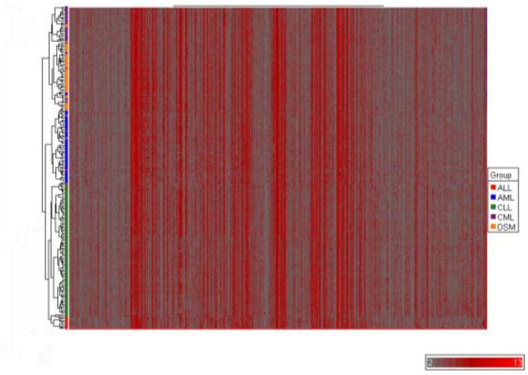


Fig. 2. Classification obtained.

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 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.

The next step, for future works, consists of incorporating new techniques for knowledge extraction, able to provide a human expert with information about the system-generated classification by means of a set of rules that are provided to support the decision-making process.

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

References

1. Shortliffe, E.H., Cimino, J.J.: Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer. (2006)
2. Tsoka S., Ouzounis C.: Recent developments and future directions in computational genomics. FEBS Letters, Vol. 480 (1), pp. 42--48 (2000)
3. Lander E.S. et al.: Initial sequencing and analysis of the human genome. Nature, Vol. 409, 860--921 (2001)
4. Rubnitz, J.E., Hijiya, N., Zhou, Y., Hancock, M.L., Rivera, G.K., Pui C.: Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatric Blood & Cancer, Vol. 44 (2), pp. 138--141 (2005)
5. Affymetrix, GeneChip Human Exon 1.0 ST Array, http://www.affymetrix.com/products/arrays/specific/Exon.affx
6. Quackenbush J.: Computational analysis of microarray data. Nature Review Genetics, Vol. 2(6), (2001). 418-427
7. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., Lockhart, D. H.: High density synthetic oligonucleotide arrays. Nature Genetics, Vol. 21, pp. 20--24 (1999)
8. Kolodner J.: Case-Based Reasoning. Morgan Kaufmann (1993)
9. Irizarry, R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. Speed,T.P.: Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. Biostatistics, Vol. 4, pp. 249--264 (2003)
10. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, pp. 59--69 (1982)
11. Fritzke, B.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, pp. 625--632 (1995)
12. Martinetz, T.: Competitive Hebbian learning rule forms perfectly topology preserving maps. ICANN'93: International Conference on Artificial Neural Networks, pp. 427--434 (1993)
13. Martinetz, T., Schulten, K.: A neural-gas network learns topologies. T Kohonen, et al. Amsterdam: Artificial Neural Networks, pp. 397--402 (1991)
14. Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition, Vol. 34, pp. 1625--1637 (2001)
15. Gariepy, R., Pepe, W. D.: On the Level sets of a Distance Function in a Minkowski Space. Proceedings of the American Mathematical Society, Vol. 31(1), pp. 255--259 (1972)
16. Jolliffe I.: Principal Component Analysis. Second Edition. Springer Series in Statistics. (2002)
17. Riverola F., Díaz F., Corchado J. M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence, Vol. 22, pp 254--268 (2006)
18. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. (1990)
19. Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Vol. 4, pp. 406--425 (1987)
20. Sneath, P. H. A., Sokal, R. R. Numerical Taxonomy. The Principles and Practice of Numerical Classication. W.H. Freeman Company, San Francisco. (1973)
21. Fix, E., Hodges, J.L., Discriminatory analysis, nonparametric discrimination consistency properties, Technical Report 4, United States Air Force, Randolph Field, TX. (1951)