Intelligent Data Analysis 14 (2010) 1–18
DOI 10.3233/IDA-2010-0417
IOS Press

CliDaPa: A new approach to combining clinical data with DNA microarrays

S. González[a], L. Guerra[a], V. Robles[a], J. M. Peña[a] and F. Famili[b]
[a]*Department of Computer Architecture, Universidad Politécnica de Madrid, Madrid, Spain*
E-mail: {sgonzalez,lguerra,vrobles,jmpena}@fi.upm.es
[b]*National Research Council of Canada, Ottawa, Canada*

Abstract. Traditionally, clinical data have been used as the only source of information to diagnose diseases. Nowadays, other types of information, such as various forms of omics data (e.g. DNA microarrays), are taken into account to improve diagnosis and even prognosis in many diseases. This paper proposes a new approach, called CliDaPa, for efficiently combining both sources of information, namely clinical data and gene expressions, in order to further improve estimations. In this approach, patients are firstly divided into different clusters (represented as a decision tree) depending on their clinical information. Thus, different groups of patients with similar behaviors are identified. Each individual group can be studied and classified separately, using only gene expression data, with different supervised classification methods, such as decision trees, Bayesian networks or lazy induction learning. To validate this method, two datasets based on Breast Cancer, a high social impact disease, have been used. For the proposed approach, internal (0.632 Bootstrap) and external validations have been carried out. Results have shown improvements in accuracy in the internal and external validation compared with the standard methods with clinical data and gene expression data separately. Thus, the CliDaPa algorithm fulfills our proposed objectives.

Keywords: Breast cancer, DNA microarray, clinical, data mining, clinical tree

1. Introduction

Nowadays, there are different different types of information to use in prognostic and diagnostic medicine. Some years ago, the only data came from clinical analyses (laboratory analysis, patient medical records and so on), which is still the basis of everyday use. Nevertheless, in the last decade DNA microarrays appeared and have since become an important aspect of research where they could play an important role in tomorrowś medicine. A DNA is a large set of hybridized DNA molecules sorted onto a solid (silicon or plastic) surface, called *biochip*. These types of experiments allow relative levels of mRNA abundance to be determined in a set of tissues or cell populations for thousand of genes simultaneously. A complete review of the methods used in the processing and analysis of gene expression data generated using DNA microarrays is defined in [15]. In [17] technical aspects and several algorithms, such as cluster analysis, gene or protein function prediction, and principle component analysis, used in standard microarray data analyses are described. Other technical applications of microarray technology, such as disease prognosis and diagnosis, drug discovery, toxicology, aging, and mental illness, are described in [30], whereas a proof survey of computational tools to analyse these data can be seen in [26].

Compared to the traditional uses of clinical information, a new approach based on clinical and DNA microarray data is proposed. Because clinical data have been the only source of information for studying diseases for several years (before the concept of microarray), it is expected that this data will be of relevant importance. In fact, all medical conclusions and experiences in all these years using clinical information must be used in an a priori decision, to help in microarray analysis conclusions. Thus, a new combined data analysis approach, called CliDaPa, is presented. This approach is based on use of clinical information a priori, categorizing and classifying different evolution of patients, to improve the microarray analysis. This categorization could be represented using a tree. This tree can represent both different categories and relationships with different clinical data.

In this paper, the CliDaPa algorithm has been applied on Breast Cancer disease data. Focusing on this application field, around 30 people die of cancer every minute. It is a problem of high social impact that must be solved as quickly as possible. For instance, breast tumor is one of the most frequent types of cancer in the world. About 135,000 new breast tumors are diagnosed and 58,000 deaths occur because of this tumor in Europe every year. In the United States, 215,900 cases are diagnosed and it causes 40,110 deaths last year (2008). In the whole world, more than a million breast tumors have diagnosed every year and there are about 400,000 deaths. With these statistics, breast tumor constitute one of the diseases with highest impact in the world, and making progress in finding a cure for this cancer is important.

Finding a cure for this type of disease would translate into a much longer life expectancy. In the scientific field, expert researchers are devoted to the study of possible solutions to these types of diseases, using all known techniques such as clinical information, experience, proteomic and genomic information, etc. Of course, a significant amount of computer resources and approaches are needed in the work routine of an expert biologist while studying disease patterns. Among the many contributions made by information technology biologists (*Bioinformatitians*), the techniques of automatic data analysis (*Computational Biology*) is one of the most promising.

The structure of the paper is as follows: The next section presents the DNA Microarray technology. Section 3 analyzes state of art of breast cancer research. Section 4 describes first CliDaPa algorithm developed. In the same section, standard methods using these data are presented too. Section 5 describes datasets used in this paper. In Section 5, tables with outcomes in internal validation are presented. An automatic complex external validation is also described in this section. Finally, all conclusions and acknowledgments are presented in the last Sections (6 and 7).

2. DNA microarray technology

DNA microarrays [15,20,26,27,30] are a relatively new and complex technology used in molecular biology and medicine. Microarrays present unique opportunities in analyzing gene expression and regulation in an overall cellular context. This technology has been applied in diverse areas ranging from genetic and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology and neurobiology. Microarray technology is the most widely used technology for the large-scale analysis of gene expression because it provides a simultaneous study of thousands of genes in a single experiment.

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides (short molecules consisting of several linked nucleotides, between 10 and 60, chained together and attached by covalent bonds), called Expressed Sequence Tags (ESTs), each containing several molecules of a specific DNA sequence. This can be a short section of a gene or other DNA element.

2.1. Steps in the DNA microarray experiment

There are several steps [20,27] in the design and implementation of a DNA microarray experiment (Fig. 1). Many strategies have been investigated in each of these steps.

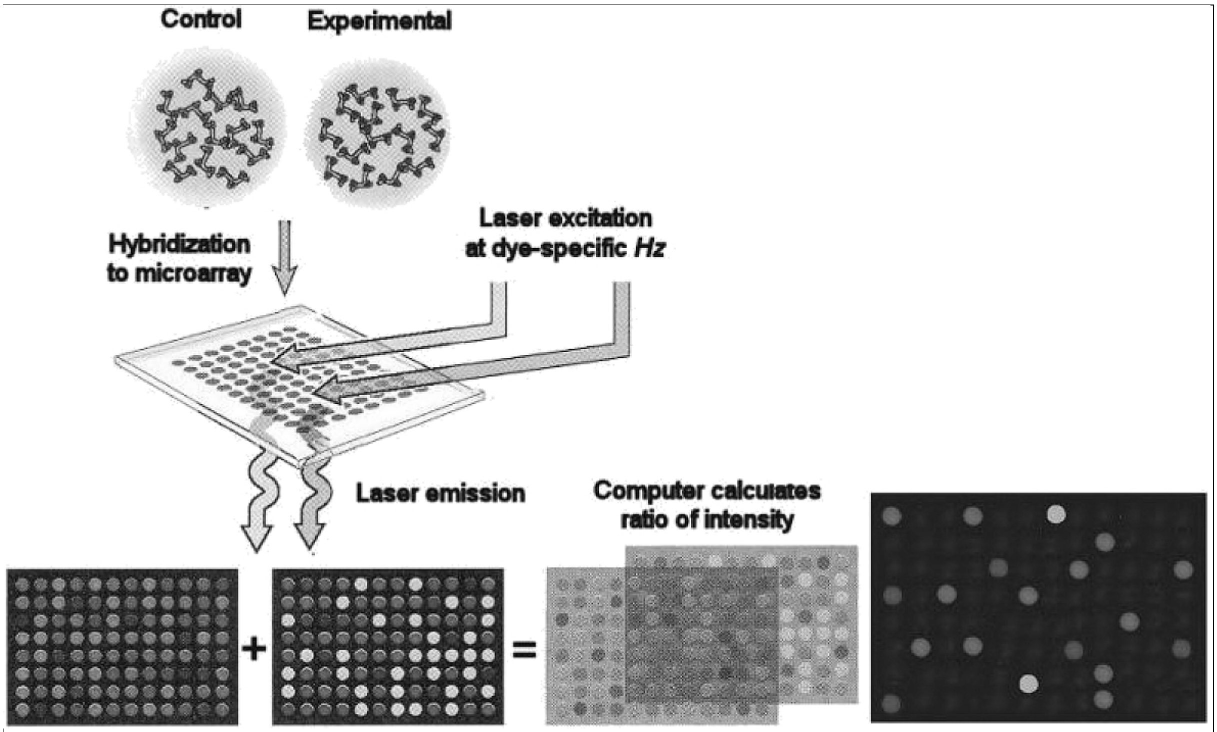

Control
Experimental
Hybridization to microarray
Laser excitation at dye-specific Hz
Laser emission
Computer calculates ratio of intensity
+
=

Fig. 1. Steps of DNA Microarray experiment. Image from Gibson & Muse 2002.

Probe First of all, the sample is obtained. The DNA type (cDNA/oligo with known identity) and the organism must be chosen in this step.

Chip manufacture The probes are placed on a surface. In standard microarrays, the information is attached to a solid surface by a covalent bond. The solid surface can be glass or silicon, in which case they are commonly known as *gene chip* or *biochip*. Here, several techniques have been used: Photolithography, pipette, drop-touch, piezoelectric (ink-jet), etc.

Sample preparation In this step the samples have been prepared. cDNA transcripts are prepared and labelled with a red fluorescent dye. A control library is constructed from an untreated source and labelled with a different fluroescent green dye.

Assay All information is hybridized (Fig. 2). Hybridization [39] is the process of combining single-stranded nucleic acids into a single molecule to the microarray.

Redaout Dual-channel laser excitation excites the corresponding dye, whose fluorescence is proportional to the degree of hybridization that has occurred. Relative gene expression is measured as the ratio of the two fluorescences: *up-regulation* of the experimental transcriptome relative to the control will be visualized as a red pseudo-colour, *down-regulation* shows as green, and constitutive expression as a neutral black. The intensity of color is proportional to the expression differential.

Informatics In this final step, new information and values are obtained from the fluorescence intensities using different computer techniques such as Robotics control, image processing, DBMS, bioinformatics, data mining and visualization.

Nowadays, there are companies that create tools for analyzing complex genetic information such as DNA microarrays. Companies such as Affymetrix [4], Celera, Gene Logic, Xenometrix, . . . have built

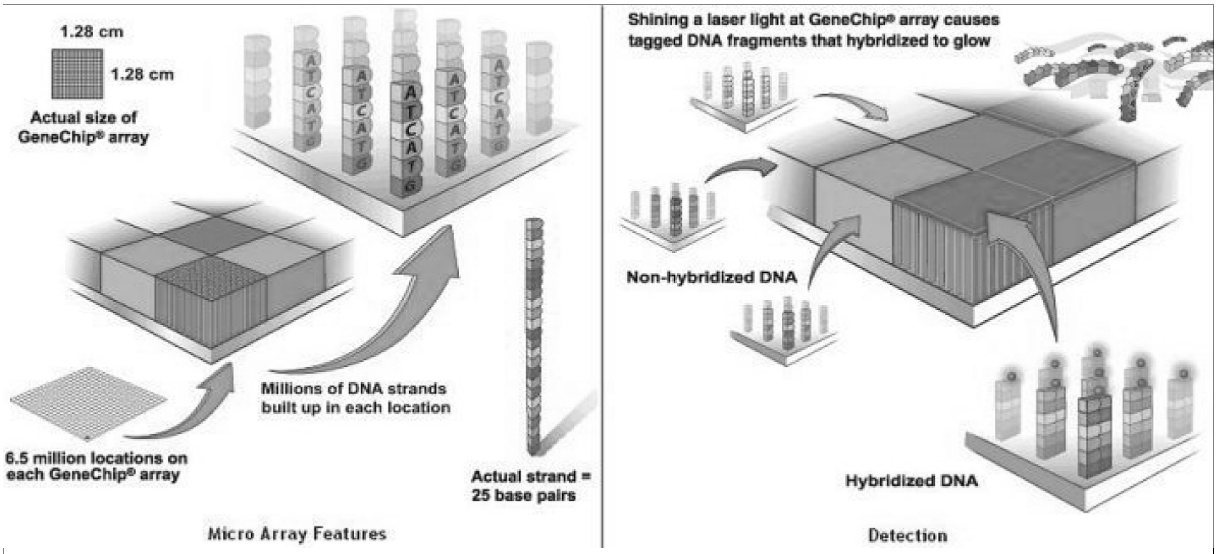

1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
Millions of DNA strands built up in each location
6.5 million locations on each GeneChip® array
Actual strand = 25 base pairs
Micro Array Features
Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA
Detection

Fig. 2. Hybridization process. Image from http://universe-review.ca/.

commercial platforms to carry out microarray experiments. Each platform obtains results using different methods (such as as Fluorescence, Mass spectrometry, Radioisotope, etc.) at each step of the microarray experiment. The use of a particular platform determines the type of experimental design possible, the type of normalization, etc.

2.2. Normalization and transformation

After the hybridization and microarray image processing from which we obtain Cy5 and Cy3 fluorescency intensities, it is needed to normalize [25,31,40] the data from each of the two scanned channels. There can be differences in labelling and detection efficiencies for the fluorescent labels and differences in the quantity of the initial values from the two samples examined in the assay. These problems can cause a shift in the average ratio of the fluorescence intensities, so they must be re-scaled before an experiment can be properly analysed. The normalization factor is used to adjust the data to compensate for experimental variability and to balance the fluorescence signals from the two samples.

There are many approaches for normalizing the gene expression data. Some, such as total intensity normalization, are based on the assumption that the quantity of the initial RNA is the same for both labelled samples, so that consequently the total integrated intensity computed for all the elements in the array should be the same in both channels. Under this assumption, a normalization factor can be calculated and used to re-scale the intensity for each gene in the array. In addition to total intensity normalization, there are a number of alternative approaches for normalizing expressions, including linear regression analysis, log centering, rank invariant methods and Chen's ratio statistics (normalization using ratio-statistics), among others [25]. However, none of these approaches takes into account systematic biases that may appear in the data: a dependence between intensity and ratio expression. Locally weighted linear regression (LOWESS) analysis [2], the most commonly used normalization method in DNA microarray experiments, can remove this dependency.

3. DNA & clinical data for breast cancer

All cancers begin in cells, the elements that generate tissues, which generate the organs of the body. Normally, cells grow and divide to form new cells. When cells grow old they then die and new cells take their place. Sometimes, this process goes wrong: creating new cells when the body does not need them, or remaining old cells do not die when they should. These *extra* cells form a mass of tissue called *tumor*. Tumors can be benign or malignant. Breast cancer is a disease in which a malignant tumor is formed in the tissues of the breast.[1]

This disease has been studied for over 35 years. A great number of different types of studies have been carried out on breast cancer. Nowadays, the most important research is classified as clinical assays, genetic and proteomic analysis.

Regarding clinical assays, there are several lines of research [24,28,29] that has resulted in some interesting findings. For example, in NIH [10] clinical data have been classified in three categories: a) patient characteristics that are independent of the disease; b) disease characteristics (such as tumor size); c) biomarkers, measurable parameters in tissues cells or fluids (estrogen receptor status). And all of them are useful for prognosis and diagnosis, but are not as accurate as they could be. In St. Gallen [6] it is stated that the presence or absence of estrogen and progesterone receptors are powerful markers to predict responsiveness.

As commented previously, a large number of cases of breast cancer are produced by hereditary transmision so it is interesting to bear in mind that gene and protein expressions can have a priori information not visible for clinical assays. Thus, in genetic and proteomic analysis, researchers are trying to find out more about breast cancer and its evolution using gene and protein expressions. Focusing on microarray technology in breast cancer [3], there is a lot of research being carried out to study different scopes of breast cancer, For example [13], report the analysis and identification of genes that allow to distinguish mutations of breast cancer (BRCA1 and BRCA2). Other research [14] tried to obtain meta-information and relationships between genes and their behavior in the disease [7] used DNA to support the idea that metastatic capacity might be an inherent feature of breast tumors.

Nevertheless, the main research group using microarray technology applies machine learning, that is supervised disease classification, non-supervised classification (genes classification) [34] and gene (feature) selection [18] has reviewed machine learning methods used, not only in breast tumors, but in bioinformatics in general, such as supervised classification, clustering and probabilistic graphical models for knowledge discovery, as well as deterministic and stochastic heuristics for optimization. Applied to breast tumor [41], presented a Bayesian approach to gene selection and classification using a logistic regression model. Others well known research improved disease prediction in young patients with breast cancer [37,38]. [12] has used DNA for gene expression monitoring and suggesting a general strategy for discovering and predicting types of cancer.

Although genetic and proteomic research has shown that breast cancer knowledge has improved [9], proposed to show us that clinical data could perform similarly to gene data if they are used in a well-trained neural network.

Other research has proposed to merge clinical assays and genetic analysis, that is to group genomic and clinical information [21,35]. For instance, in [1] subtypes of cancer have been indentified using gene expression and clinical data [32] tried to classify breast carcinomas based on variations in gene expression pattern and to correlate tumor characteristics to clinical outcome [8] suggested that due to

[1] To know more about this kind of cancer, please visit http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.jsp.

Algorithm 1 Preprocess

```
for each (continuous clinical data)
  for $k = 2$ to 6 do
    K-means
  end for
end for
```

the complexity of breast tumor prognosis, clinical and gene data could be both used trying to find complementary information. In this way [11,19,33,35], are studies which combine both types of data, either using Bayesian networks or logistic regression.

Following this line of research and taking into account that clinical data were the only source of information until some years ago, for which it should have some important or complementary information, a new combined method using clinical and genetic data is proposed. In this approach, an initial dataset is divided into smaller datasets, through clustering instances based on clinical data. With this divide-and-conquer technique, information is simplified until the best possible outcome is obtained by filfilling some restrictions. All this research is focused on two breast cancer datasets used in [37,38] whose outcome is whether patients relapse or not.

4. CliDaPa

In this section, the Clinical Data Partition (CliDaPa) algorithm is described. The idea of this paper is to identify differences between the traditional use of clinical and/or genetic information and CliDaPa algorithm. So firstly, the standard methods of data mining is presented for later introduction into the CliDaPa approach.

4.1. Traditional use

Three experiments have been carried out in this paper associated to traditional use:

- Working with clinical data only
- Working with DNA microarray data only
- Merging both clinical and DNA microarray data.

All experiments try to estimate the outcome (relapse or not relapse) using different supervised learning methods, such as Naive Bayes, K nearest neighbour and decision tree algorithms [18,22]. Finally, the accuracy is obtained and related to the accuracy, comparing it with real outcome.

4.2. Clinical data partition

The algorithm, called CliDaPa (Clinical Data Partition), consists of two steps: A data preprocessing step (Algorithm 1) and a greedy algorithm (Algorithm 2).

4.2.1. Preprocess

To proceed with the greedy algorithm, it is necessary for all clinical data to be discrete values. So if there are continuous values they must be discretized using any mechanism. In this case, a K-means algorithm [16] is used to discretize data. Five new variables are obtained (Fig. 3) from each continuous feature using K-means. These new variables describe from 2 to 6 clusters or partitions of the continuous variable.

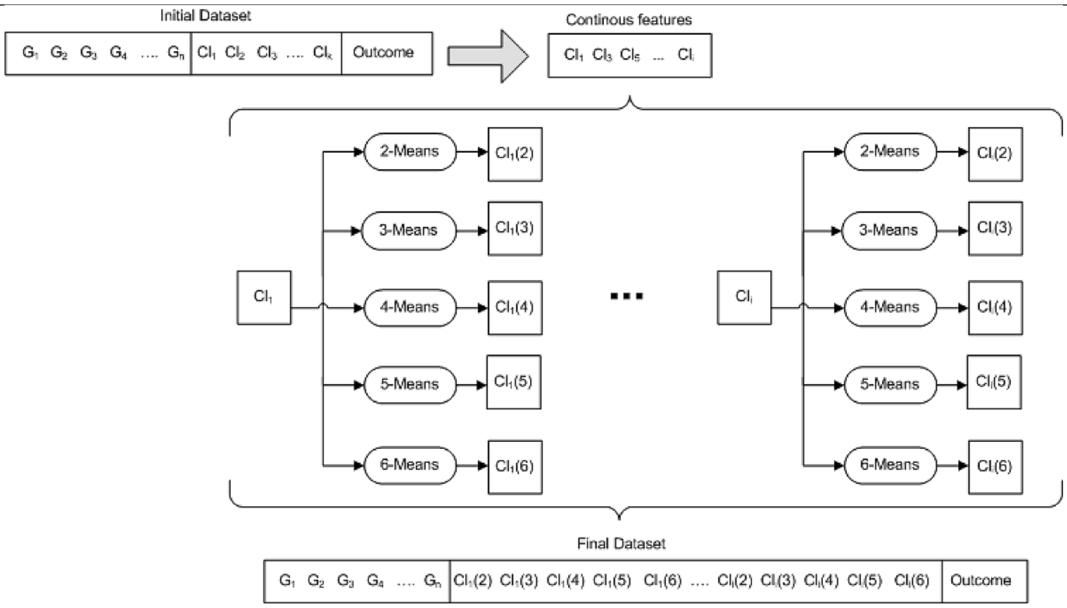


Fig. 3. Preprocess algorithm.

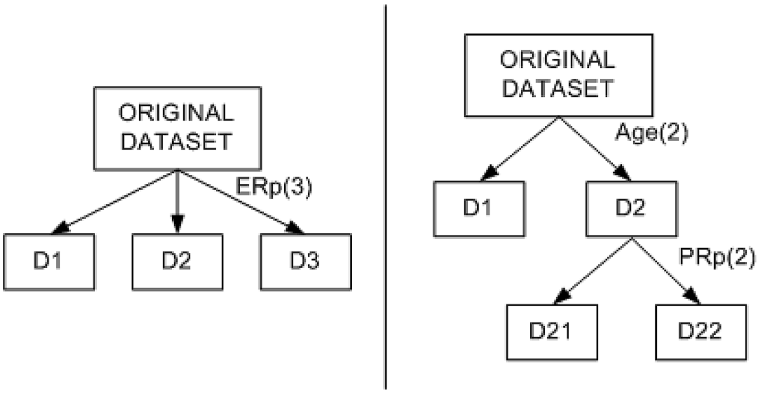


Fig. 4. Examples of clinical tree.

4.2.2. Greedy algorithm

Once the data is preprocessed, the second step, a greedy algorithm, starts. The algorithm is an iterative process consisting of partitioning data using clinical data, that is, to split this initial dataset into smaller datasets based on clinical data. In each iteration, all data and sub-datasets are partitioned for each clinical variable. sub-datasets created from one clinical variable, form leaves of the tree that is being created. Figure 4 shows two examples of decision trees with different leaves and children.

Each leaf is evaluated using an internal 0.632 Bootstrap algorithm [5] with a supervised classification algorithm [22](Naive Bayes, C4.5 or KNN). Parent leaf accuracy is obtained by weighting the results of all its children, using an accuracy aggregation (with a proportional weight).

$$Acc_{weighted} = \frac{Acc_1 x Instances_1 + Acc_2 x Instances_2 + \ldots + Acc_n x Instances_n}{Instances_1 + Instances_2 + \ldots + Instances_n}$$

Algorithm 2 Greedy iteration i

```
for all clinical_data as cd
    precandidate = (divide(dataset_i, cd))
    precandidates.add(precandidate)
end for
candidate = select_best(precandidates)
if (candidate.getAccWeighted() > dataset_i.getAccNotPartitioned()) then
    clinical_data.erase_clinical_data(candidate.getClinicalData())
    for all candidate.leaves as leaf do
        greedy_clidapa(leaf)
    end for
end if
```

Overfitting problems can appear in the classification learning when any dataset d_i is divided into N sub-datasets using any clinical data and the number of instances obtained in there is small. For that reason, each dataset has to fulfil a condition to be evaluated: the number of instances has to be greater than a certain *threshold* β.

The iterative process continues while the aggregated accuracy of every given partition improves the accuracy from the unpartitioned leaf. Going into more detail, in each iteration, once $Acc_{weighted}$ is calculated, it is compared with the accuracy when the dataset is not partitioned ($Acc_{notpartitioned}$). If $Acc_{weighted} > Acc_{notpartitioned}$, then these leaves will be effective. Once this partition is choosen as *winner*, the clinical variable selected is removed from the set of clinical information to the next iteration. If the clinical variable is created in the preprocess (using k-means), all the variables created using the same original continuous clinical data will be removed too.

The entire process (Fig. 5) is repeated until all the created datasets have fewer instances than the required threshold β or there is no clinical data left to divide. When the greedy process finishes, the tree is created and the final accuracy is obtained by carrying out a new weight that has the accuracy values of all the best leaves selected from each iteration of the algorithm.

From the CliDaPa algorithm two results are interesting. The final accuracy obtained, which is an internal validation value, and the tree obtained (with the clinical data used to get it), which will be used later to carry out a external validation.

4.3. Overfitting

When the number of instances is very small, overfitting cases could be appear. Internally, the CliDaPa algorithm can work without any problem, obtaining a very good accuracy in the internal Bootstrap. However, when the CliDaPa tree is externally validated, a very poor accuracy is obtained. To solve this problem, a very simple improvement can be proposed: To make a set of CliDaPa trees and select the tree with the lowest accuracy. The rationale behind this idea is as follows:

- with many iterations of Bootstrap (500 or more), the accuracy variance is low (that is, all results are similar), so although the tree with the lowest accuracy is selected, this tree presents overfitting (because it is one of the best results).
- with a low number of iterations (50 or less), the accuracy variance is very high (different results). Selecting the tree with the lowest accuracy may or may not be a good decision, because this can be created using the real good attributes or not (too much variance).

Thus, this solution is not the best one. However, related to the tree selection approach and associated to an external K-fold validation, a new approach is proposed. In each fold an overfitted (only in this fold)

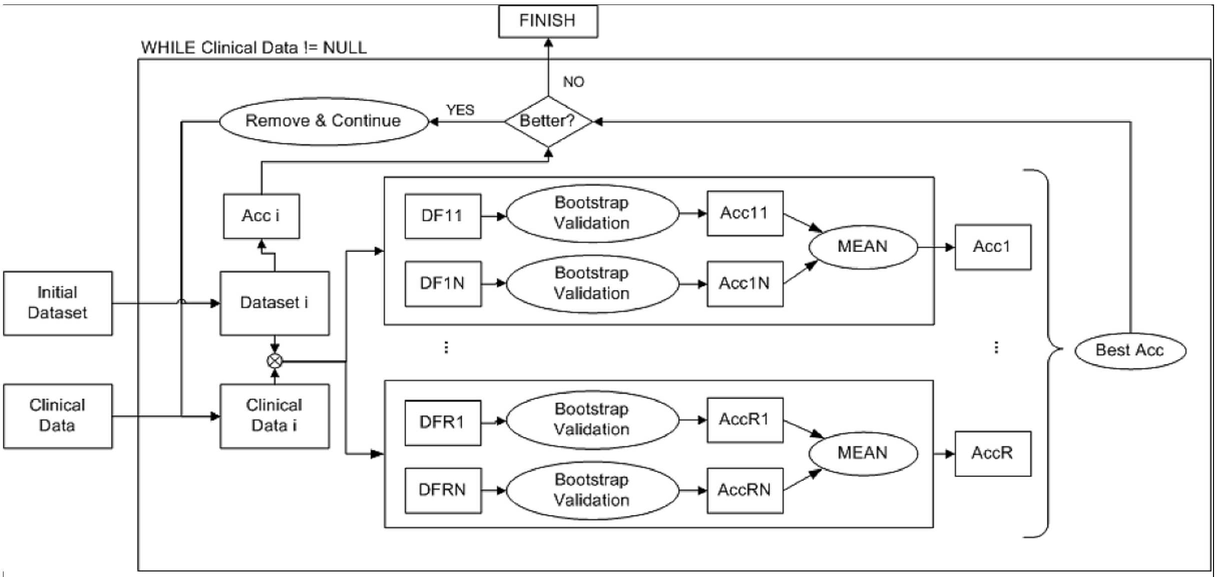


Fig. 5. CliDaPa diagram.

CliDaPa tree is obtained. Thus K trees are obtained. Next, the tree structure with the lowest accuracy is selected to be applied to all the folds. So in each fold, data is already divided and the selected tree structure is defined. With this approach, the overfitting generated is lower. Futher information about CliDaPa With Overfitting Attenuation can be found in the *Validation* section.

5. Datasets

The following two datasets on Breast Cancer with clinical data are presented in this section.

5.1. Van't Veer dataset

The Van't Veer dataset [38] on Breast Cancer[2] has been considered to validate our approach. Van't Veer results have been approved by the FDA (Food and Drug Administration) and were applied in a genomic profiling test, called MammaPrint, that predicts whether patients will suffer breast cancer relapse or not. The data is divided into two groups, learning and validation instances. The training data consist of 78 patients, 34 of whom are patients that developed distance metastases within 5 years (poor prognosis). The rest of the dataset (44 patients) are the ones who remained healthy from the disease after their initial diagnosis for an interval of 5 years (good prognosis). The second group of patients (validation dataset) consists of 19 patients, 12 patients with poor prognosis and 7 with good prognosis. DNA microarray analysis was used to determine the mRNA expression levels of approximately 24,500 genes for each patient. All the tumors were hybridized against a reference pool made by pooling equal amounts of RNA from each patient.

[2] available at http://www.rii.com/publications/2002/vantveer.html.

5.1.1. Preprocessing

Obviously real data have a lot of redundancy, as well as incorrect or missing values, depending on some factors. So, as first step, we carry out some pre-processing in order to clean up and prepare the data. We also discard variables with low internal variance or low Pearson correlation with outcome.

Several preprocessing algorithms have been carried out through the training data. Firstly, we have discarded genes that are replicated. Next, we have discarded patients that had more than 80% of missing gene values. All data have been background corrected, normalized and log-transformed using Lowess Normalization [25]. Missing values were estimated using a 15-weighted nearest neighbor algorithm [36] (kNN Impute).

5.1.2. Biomarker selection

The objective of this paper is not to obtain a feature selection of 24,500 genes expressions, but to demostrate the improvement when clinical data helps to organizate microarray data and to obtain better accuracy in the outcome estimation. Thus, the microarray data is filtered to the 70 Van't Veer [38] selected genes (accepted by the FDA as breast cancer biomarkers).

5.2. Van de Vijver dataset

The Van de Vijver dataset [37] consists of tumors from a series of 295 consecutive women with breast cancer from the fresh frozen tissue bank of the Netherlands Cancer Institute according to the following criteria:

- the tumor was invasive breast carcinoma that was less than 5 cm in diameter
- lymph nodes were tumor-negative,
- the age was 52 years or younger
- the calendar year of diagnosis was between 1984 and 1995
- there was no previous history of cancer (except nonmelanoma skin cancer).

Among the 295 patients, 151 had lymph node negative disease (pN0) and 144 had lymph node positive disease (pN+). 10 out of the 151 patients who had pN0 and 120 out of the 144 who had pN+ had received adjuvant systemic therapy consisting of chemotherapy (90), hormonal therapy (20), or both (20). Of all 295 patients, 101 women have good prognosis in breast cancer and 194 have poor prognosis. Follow-up information was extracted from the medical registry of the Netherlands Cancer Institute.

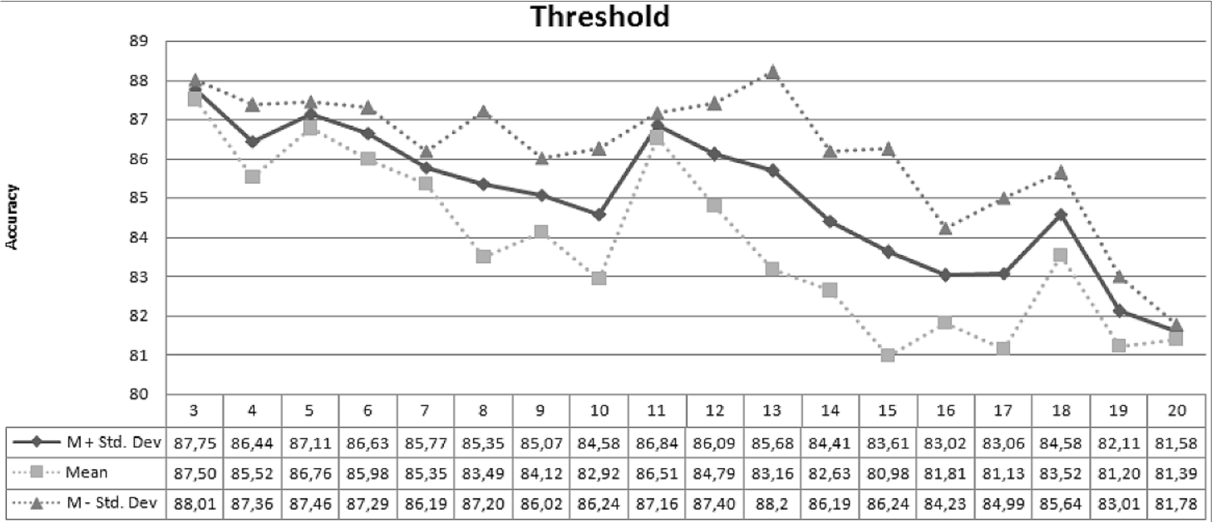
6. Results

This section presents the study of the threshold β and an external validation mechanism used in all the executions. Next, all executions are presented and compared.

6.1. Optimal number of minimal instances

To obtain all results, a threshold β is needed. Some tests have been carried out to define one or more β values. These tests are about 170 executions of CliDaPa using NaiveBayes method and varying β value from 3 to 20. NaiveBayes is selected because, in cases with curse of dimensionality, it is the method that will result with minor overfitting [23]. Looking at Fig. 6, several conclusions can be extrated:

- Executions with small β can produce overfitting.



	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M+ Std. Dev	87,75	86,44	87,11	86,63	85,77	85,35	85,07	84,58	86,84	86,09	85,68	84,41	83,61	83,02	83,06	84,58	82,11	81,58
Mean	87,50	85,52	86,76	85,98	85,35	83,49	84,12	82,92	86,51	84,79	83,16	82,63	80,98	81,81	81,13	83,52	81,20	81,39
M- Std. Dev	88,01	87,36	87,46	87,29	86,19	87,20	86,02	86,24	87,16	87,40	88,2	86,19	86,24	84,23	84,99	85,64	83,01	81,78

Fig. 6. Threshold evaluation.

- Executions with high β obtain very simple but unaccurated trees.
- There are two significant β values, 5 and 11. Both have very small standard deviation and a high accuracy value.
- The first one has overfitting, which can be reduced using the approach defined in Section 4.3.
- The second one remove the possibility of divisions using any attribute, even more if the data have a low number of instances.

Thus, value 5 is selected as the threshold β in all CliDaPa executions.

6.2. Validation

The validation process is the essencial part of an algorithm evaluation that gives us more reliability of the algorithm results. In this research two different validations have been carried out. The first one, called internal validation, is the 0.632 Bootstrap process. The second one, called external validation, is based on an MxN fold cross validation, where N is the number of folds and M is the number of times all the process is carried out.

All approaches presented in this paper (existing methods and CliDaPa) have an external validation based on a 10 × 5 fold cross validation. CliDaPa executions also have 500 iterations of Bootstrap internal validation. Figure 7 shows us the external validation divided into 5 phases: File Distribution (dataset is divided into N folds, and each fold has learn and test sub-datasets), CliDaPa algorithm (tree is obtained here), the division of test sub-dataset defines the CliDaPa tree, learn and validate with a specific learning model, and finally obtain final average of the accuracies.

The external validation process has few changes when overfitting problems appear. First, all CliDaPa trees are obtained. Next, the tree with the lowest accuracy (obtained in internal Bootstrap validation) is selected as the tree which is applied to all the test sub-datasets of all folds. Finally, all sub-datasets are divided as defines the selected tree and continues with the rest of phases normally. This modified process is called CliDaPa WOA (With Overfitting Attenuation).

Other possible characteristic in phase 1 of the external validation (fold division) is the fold division. This could be done randomly (selecting instances randomly) or using any balanced technique, that is,

Table 1
External validation of Vant' Veer data using standard methods

Division	Type	Method	Times	Folds	Mean Acc	Std Dev Acc
Balanced	Genes + Clinical	KNN	10	5	64.89	2.70
		C4.5			63.42	3.89
		NB			79.76	1.06
	Genes	KNN			75.63	2.04
		C4.5			63.02	2.88
		NB			80.10	0.99
	Clinical	KNN			63.33	2.56
		C4.5			62.04	4.12
		NB			71.83	1.96
Not Balanced	Genes + Clinical	KNN	10	5	66.01	1.77
		C4.5			62.38	5.66
		NB			79.66	1.17
	Genes	KNN			76.03	2.57
		C4.5			63.32	2.70
		NB			79.45	0.58
	Clinical	KNN			61.98	2.50
		C4.5			62.55	3.74
		NB			72.55	3.03

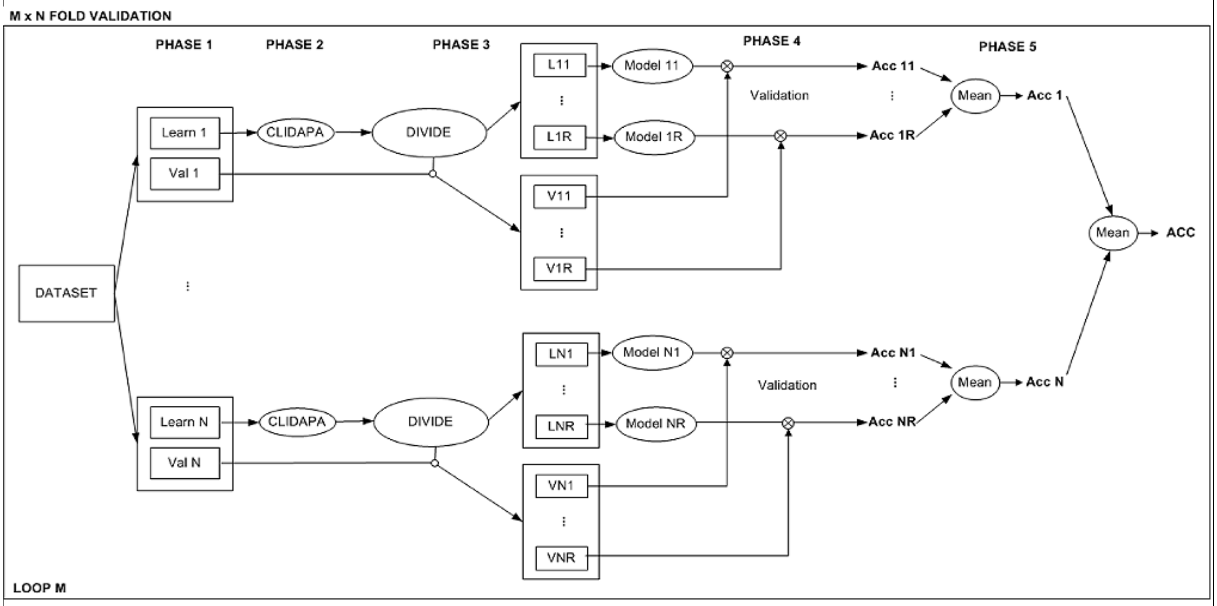


Fig. 7. External validation diagram.

when the dataset is divided into sub-datasets, the number of instances with a specific outcome value in each sub-dataset must be well-balanced.

6.3. Standard methods

Next, the following standard methods, using external 10 × 5 fold validation, and *balanced* and *not balanced* divisions, are presented. Table 1 presents the accuracy results of the Vant Veer dataset, while Table 2 shows us the accuracy results of the Van der Vijver dataset.

Table 2
External validation of Van der Vijver data in Traditional uses

Division	Type	Method	Times	Folds	Mean Acc	Std Dev Acc
Balanced	Genes + Clinical	KNN	10	5	66.27	1.21
		C4.5			74.64	1.06
		NB			70.58	0.62
	Genes	KNN			63.73	2.52
		C4.5			62.24	1.92
		NB			67.32	0.63
	Clinical	KNN			84.03	1.07
		C4.5			84.41	0.98
		NB			77.52	0.74
Not Balanced	Genes + Clinical	KNN	10	5	66.41	1.33
		C4.5			75.93	2.22
		NB			70.98	0.46
	Genes	KNN			64.24	0.98
		C4.5			61.15	2.21
		NB			67.15	0.67
	Clinical	KNN			83.90	1.44
		C4.5			84.10	1.28
		NB			77.66	0.41

Table 3
External validation of Vant Veer data using CliDaPa

Division	Type	Iterations	Times	Folds	Threshold	Method	Mean Acc	Std Dev Acc
Balanced	Regular	500	10	5	5	KNN	67.73	4.40
						C4.5	60.75	4.93
						NB	75.59	4.47
	WOA	500	10	5	5	KNN	76.31	2.69
						C4.5	67.28	4.76
						NB	81.11	1.99
Not Balanced	Regular	500	10	5	5	KNN	69.28	4.67
						C4.5	59.24	4.39
						NB	75.03	3.14
	WOA	500	10	5	5	KNN	73.41	4.16
						C4.5	63.26	6.05
						NB	81.16	2.79

6.4. CliDaPa

CliDaPa executions, using both external 10 × 5 fold validations types (Standard and WO), and *balanced* and *not balance* divisions, are presented. Table 3 presents the accuracy results of the Vant Veer dataset, while Table 4 shows us accuracy values of the Van der Vijver dataset.

7. Discussion

In both datasets, using the CliDaPa algorithm the accuracy is improved comparing the standard methods. In the Vant Veer dataset, the best accuracy with standard methods is found by using NaiveBayes and Balanced division (80.10%). With CliDaPa without overfitting (WO) and the same configuration (NaiveBayes and Balanced division) the accuracy obtained is 81.11%. In the Van de Vijver dataset, the best accuracy with traditional uses is found by using C4.5 algorithm and Balanced division 84.41%.

Table 4
External validation of Van der Vijver data using CliDaPa

Division	Type	Iterations	Times	Folds	Threshold	Method	Mean Acc	Std Dev Acc
Balanced	Regular	500	10	5	5	KNN	84.85	1.53
						C4.5	82.54	0.77
						NB	78.71	0.81
	WOA	500	10	5	5	KNN	83.66	2.44
						C4.5	85.66	0.90
						NB	79.32	1.00
Not Balanced	Regular	500	10	5	5	KNN	84.17	1.45
						C4.5	82.88	2.13
						NB	78.71	1.51
	WOA	500	10	5	5	KNN	84.47	0.78
						C4.5	82.98	2.14
						NB	78.44	1.47

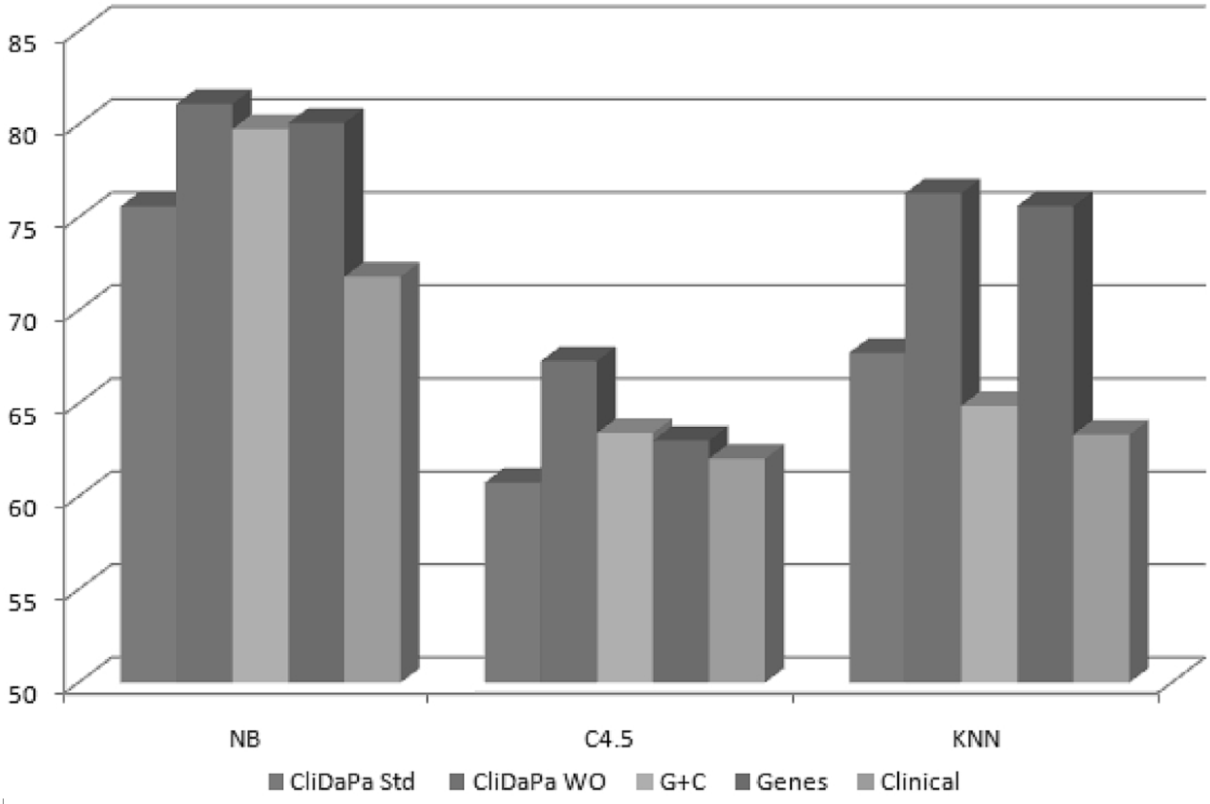


Fig. 8. Accuracies in the Vant Veer dataset.

With both CliDaPa execution (stantard and without overfitting) and the same configuration, the accuracy obtained is 84.85% and 85.66%, respectively. With other configurations on both datasets, CliDaPa algorithm also improves the accuracy. Figure 8 shows accuracies of all executions in Vant Veer dataset, while Fig. 9 presents accuracies of all executions in Van der Vijver dataset. Notice that in both figures the Y-scale has been modified from 50% to 90% of accuracy.

The CliDaPa algorithm without overfitting (WO) is useful when datasets have a low number of

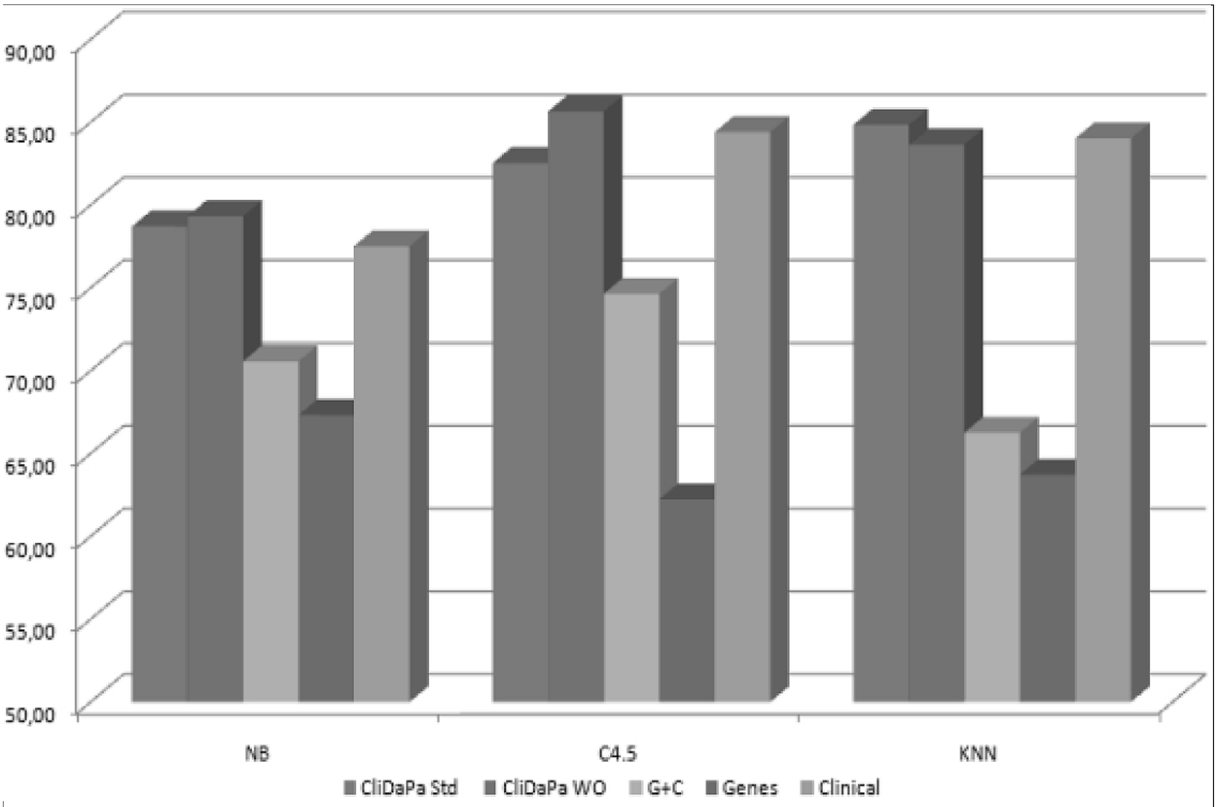


Fig. 9. Accuracies in the Van der Vijver dataset.

instances. Related to the external validation, the dataset is divided into 5 folds and each fold has 4/5 of the original dataset. Like Vant Veer the dataset has 96 instances, thus each fold has 76 instances. If a Bootstrap mechanishm is used with each fold, each Bootstrap iteration will have datasets of 51 instances to learn. This could lead to an overfitting problem. Thus, by applying CliDaPa WOA into the Vant Veer dataset, all accuracy results have been improved compared to the standard CliDaPa. With the other dataset, the overfitting problem dissapears because it has 295 instances. Results of both algorithms in the Van der Vijver dataset demostrate this statement.

Executions with balanced division generated similar results facing executions with not balanced partitioning in all possible configurations. Thus it is stated that both standard methods and CliDaPa work perfectly regardless of which kind of partitioning is carried out.

7.1. Conclusions

Results obtained in this research demonstrate that clinical data contains some very useful information which must not be ignored. It is the main conclusion obtained from the research reported in this paper, and it ought to be taken into account in order to continue improving results. Despite new types of knowledge discovered from the analysis of microarray data (such as DNA microarrays), it is important to take out all the useful information of each disposable data instead of replacing the older data.

Focusing on CliDaPa algorithm, allows us to discretize automatictly clinical real-value features. Futhermore, this algorithm selects and filters those features that can help in the outcome classification, that is, those features that have high entropy with the outcome. However, CliDaPa can fall into a typical overfitting problem with datasets with low number of instances. But, in the same research, a modification in the algorithm is proposed to remove some overfitting.

The CliDaPa algorithm can help us to understand the disease behaviour and to group common behaviour between patients (instances). The final results of the algorithm is not only the accuracy, but also the clinical tree itself, that represents the partition of patients depending on the clinical features selected. In fact, this algorithm divides and groups the dataset into smaller ones that reprensent each one a common behaviour. These common behaviours help us to better learn the outcome. Eventually the results obtained have to be validated by cliniciians.

As an interesting conclusion of this research, the challenges that one would face when combining various forms of data, such as combining microarray and clinical, are solved using the CliDaPa algorithm. This can be applied, not only to medical and biological fields such is SNPs, but all scientific fields that have real and discrete data came from different sources.

8. Acknowledgments

The authors are grateful to Belen Cañón, Pedro Larrañaga, the Blue Brain Project Team and the IT-NRC Team for their technical assistance. They also thankfully acknowledge the computer resources, technical expertise and assistance provided by the Centro de Supercomputaci ón y Visualización de Madrid (CeSViMa) and the Spanish Supercomputing Network.

References

[1] E. Bair and R. Tibshirani, Semi-supervised methods to predict patient survival from gene expression data, *PLoS Biol* **2**(4) (April 2004).

[2] William S. Cleveland, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**(368) (1979), 829–836.

[3] C. Cooper, Applications of microarray technology in breast cancer research, *Breast Cancer Res* **3** (3) (2001), 158–175.

[4] D.D. Dalma-Weiszhausz, J. Warrington, E.Y. Tanimoto and C.G. Miyada, The affymetrix genechip platform: an overview, *Methods in enzymology* **410** (2006), 3–28.

[5] B. Efron and R. Tibshirani, Improvements on cross-validation: The 0.632 bootstrap method, *American Statistical Assotiation* (92) (1997), 548–560.

[6] A. Goldhirsch et al., Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer, *J Clin Oncol* (21) (2003).

[7] B. Weigelt et al., Breast cancer metastasis: markers and models, *Nat Rev Cancer* (5) (2005), 591–602.

[8] J. Brenton et al., Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J Clin Oncol* (23) (2005), 7350–7360.

[9] P. Edén et al., 'good old' clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers, *Eur J Cancer* (40) (2000), 1837–1841.

[10] P. Eifel et al., National institutes of health consensus development conference statement: adjuvan therapy for breast cancer, *J Natl Cancer Inst* (93) (2000).

[11] O. Gevaert, F.D. Smet, D. Timmerman, Y. Moreau and B.D. Moor, Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks, *Bioinformatics* **22**(14) (July 2006).

[12] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield and E. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* (286) (1999), 531–537.

[13] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O.P. Kallioniemi, B. Wilfond, A. Borg and J. Trent, Gene-expression profiles in hereditary breast cancer, *N Engl J Med* **344**(8) (February 2001), 539–548.

[14] E. Huang, S.H. Cheng, H. Dressman, J. Pittman, M.H. Tsou, C.F. Horng, A. Bild, E.S. Iversen, M. Liao, Ch. M. Chen, M. West, J.R. Nevins and A.T. Huang, Gene expression predictors of breast cancer outcomes, *Lancet* **361**(9369) (May 2003), 1590–1596.

[15] Wolfgang Huber, Anja Von Heydebreck, and Martin Vingron. Analysis of microarray gene expression data. In *in Handbook of Statistical Genetics, 2nd edn*. Wiley, 2003.

[16] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, *IEEE Trans Pattern Anal Mach Intell* **24**(7) (July 2002), 881–892.

[17] S. Knudsen, *A biologist's Guide to Analysis of DNA Microarray Data*, John Willey and Sons, 2002.

[18] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafe, A. Perez Martinez and V. Robles. Machine learning in bioinformatics, *Briefing in Bioinformatics* (7) (2006), 86–112.

[19] L. Li, Survival prediction of diffue larg-b-cell lymphoma based on both clinical and gene expression information, *Bioinformatics* **22**(4) (2006), 466–471.

[20] D.J. Lockhart and E.A. Winzeler, Genomics, gene expression and dna arrays, *Nature* **405**(6788) (June 2000), 827–836.

[21] X. Lu, X. Lu, Z.C. Wang, J.D. Iglehart, X. Zhang and A.L. Richardson, Predicting features of breast cancer with gene expression patterns, *Breast Cancer Research and Treatment* **108**(2) (March 2008), 191–201.

[22] T. Mitchell, *Machine Learning*, McGraw-Hill Education (ISE Editions), October 1997.

[23] Andrew Y. Ng, Preventing "overfitting" of cross-validation data. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 245–253, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[24] S. Paoli, G. Jurman, D. Albanese, S. Merler and C. Furlanello, Integrating gene expression profiling and clinical data, *International Journal of Approximate Reasoning* (47) (2008), 58–69.

[25] J. Quackenbush, Microarray data normalization and transformation – nature genetics.

[26] J. Quackenbush, Computational analysis of microarray data, *Nat Rev Genet* **6**(2) (June 2001), 418–427.

[27] J. Quackenbush, Computational approaches to analysis of dna microarray data, *Methods Inf Med* **45**(Suppl 1) (2006), 91–103.

[28] C. Ritz, *Comparing prognostic markers for metastases in breast cancer using artificial neural networks*, PhD thesis, Lund University, Sweden, 2003.

[29] J.A. Gómez Ruiz, J.M. Jerez Aragones, J. Muñoz Perez, and E. Alba Conejo, Un modelo para la prediccion de recidiva de pacientes operados de cancer de mama (cmo) basado en redes neuronales, *Revista Iberoamericana de Inteligencia Artificial* (11) (2000), 39–45.

[30] M. Schena, R.A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier and R.W. Davis, Microarrays: biotechnology's discovery platform for functional genomics, *Trends Biotechnol* **7**(16) (July 1998), 301–306.

[31] G.K. Smyth and T. Speed, Normalization of cdna microarray data, *Methods* **31**(4) (December 2003), 265–273.

[32] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, Per E. Lonning and A. Borresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *PNAS* **98**(19) (September 2001), 10869–10874.

[33] C. Sotiriou, So. Neo, L.M. Mcshane, E.L. Korn, P.M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A.L. Harris and E.T. Liu, Breast cancer classification and prognosis based on gene expression profiles from a population-based study, *Proceedings of the National Academy of Sciences* **100**(18) (September 2003), 10393–10398.

[34] C. Sotiriou, N.E.O. Sy, Mcshane Lm, Korn El, Long Pm, A. Jazaeri, P. Martiat, Fox Sb, Harris and Liu Et, Breast cancer classification and prognosis based on gene expression profiles from a population-based study, *Proc Natl Acad Sci* **100**(18) (2003), 10393–10398.

[35] Y. Sun, S. Goodison, J. Li, L. Liu and W. Farmerie, Improved breast cancer prognosis through the combination of clinical and genetic markers, *Bioinformatics* **23**(1) (2007), 30–37.

[36] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman, Missing value estimation methods for dna microarrays, *Bioinformatics* **17**(6) (June 2001), 520–525.

[37] M.J. van de Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend and R. Bernards, A gene-expression signature as a predictor of survival in breast cancer, *N Engl J Med* **347**(25) (December 2002), 1999–2009.

[38] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**(6871) (January 2002), 530–536.

[39] J.G. Wetmur, Dna probes: applications of the principles of nucleic acid hybridization, *Crit Rev Biochem Mol Biol* **26**(3-4) (1991), 227–259.

[40] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai and T.P. Speed, Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res* **30**(4) (February 2002).

[41] X. Zhou, K.Y. Liu and S.T. Wong, Cancer classification and prediction using logistic regression with bayesian gene selection, *J Biomed Inform* **37**(4) (August 2004), 249–259.