Personalizing Breast Cancer Patients with Heterogeneous Data

Pedro Henriques Abreu[1], Hugo Amaro[1], Daniel Castro Silva[1], Penousal Machado[1], and Miguel Henriques Abreu[2]

[1] Department of Informatics Engineering, University of Coimbra, Portugal
[2] Portuguese Institute of Oncology of Porto, Porto, Portugal

Abstract— **The prediction of overall survival in patients has an important role, especially in diseases with a high mortality rate. Encompassed in this reality, patients with oncological diseases, particularly the more frequent ones like woman breast cancer, can take advantage of a very good customization, which in some cases may even lead to a disease-free life. In order to achieve this customization, in this work a comparison between three algorithms (evolutionary, hierarchical and k-medoids) is proposed. After constructing a database with more than 800 breast cancer patients from a single oncology center with 15 clinical variables (heterogeneous data) and having 25% of the data missing, which illustrates a real clinical scenario, the algorithms were used to group similar patients into clusters. Using Tukey's HSD (Honestly Significant Difference) test, from both comparison between k-medoids and the other two approaches (evolutionary and hierarchical clustering) a statistical difference were detected ($p - value < 0.0000001$) as well as for the other comparison (evolutionary versus hierarchical clustering) – $p - value = 0.0061354$ – for a significance level of 95%.**

The future work will consist primarily in dealing with the missing data, in order to achieve better results in future prediction.

Keywords— **Women Breast Cancer, Patient Personalization, Genetic Algorithm, Clustering Algorithms.**

I. INTRODUCTION

Nowadays, Cardiovascular Diseases including Coronary Heart Disease, stroke and Heart Failure, are the main cause of death in Europe, with 4 million deaths each year (approximately 47% of all deaths) [1]. In spite of this fact, Cancer diseases rank second, presenting a very slight difference to the ones previously mentioned. In this particular context, breast cancer is the most common one in women, estimating 29% of new cases and 26% of causes of death per year in the total of all cancer cases [2].

In recent times the study of this disease has known considerable advances. The research mark in the past two decades was the discovery of HER2 (Human Epidermal Growth Factor Receptor 2), which showed that patients' treatment must be supported by a molecular understanding of breast tumors. This new marker was only detected in nearly 20% of the cases, but predicts a bad survival. The work of Slamon et al. [3] was the paradigm of this, demonstrating a survival benefit of HER2 blockage (with a drug called trastuzumab) associated with a classical chemotherapy regimen. Despite the early enthusiasm with this discovery, there have been few new prognostic markers in breast cancer after that. The gene signatures, as Mamaprint [4] [5], try to identify patients at high risk of distant recurrence following surgery, based on the analysis of many genes; however, the majority of these gene signatures is not validated for clinical practice nor cost-effective [6], and clinicians still decide based on a set of variables (patient- and tumor-dependent). Despite such advances, in 2007 only 120 articles were found relating cancer prediction/prognosis with soft-computing techniques [7]. Throughout the years, many authors have tried to predict, for instance, the overall survival of breast cancer patients [8] [9] using public datasets (e.g. SEER – Surveillance Epidemiology and End Results[1]). However, while the results are promising, some technical question are still unanswered: What will the behavior of these algorithms be in missing data contexts, such as a real clinical environment? In those scenarios, what will be the best strategy to be adopted in order to decrease the noise added by the missing data? Is this strategy influenced by the type of missing data present in the dataset? In general, these questions are solved by researchers using methods that do not attend to the nature of the data [10], which is far from the best approach in a missing data context. Encompassed in a wider project where the goal is to identify the real impact of using data mining techniques to solve different types of missing data in the prediction of breast cancer patient having overall survival as the target problem, in this project a personalization of breast cancer patients was performed using three distinct algorithms. Based on 15 variables that are available in clinical practice (linear and nominal variables) the three algorithms tried to identify the distinct patients groups using a heterogeneous distance measure. The obtained results were very promising, illustrating the broad spectrum of patients stored in the database.

The remainder of this paper is organized as follows: Section II presents a brief review of the literature, while section III outlines the methodological steps used in this project and section IV presents the collected results. Finally, in section V, the conclusions and some proposals for further studies are presented.

[1] available at `http://seer.cancer.gov/data/`

I. Lacković et al. (eds.), *The International Conference on Health Informatics*, IFMBE Proceedings 42,
DOI: 10.1007/978-3-319-03005-0_11, © Springer International Publishing Switzerland 2014

II. Literature Review

Personalizing patients is a hard topic, especially in oncological diseases, where a good customization directly influences the response of the patients to a given treatment. In consequence of that, many studies tried to identify the main characteristics of a certain cancer disease [11] [12]. Over the years, different types of studies emerged in oncology, basing their knowledge on:

1. Images to support the diagnose or prediction [13]. Using microscopic images, Keskin et al. [14] tried to classify 14 different classes (7 classes of breast cancer and 7 for liver cancer) using Support Vector Machines, achieving 98% of accuracy.
2. Genetic information to the identification of new biomarkers that can influence the prediction or prognostic for a certain cancer [15]. Combining Genetic (to select the best features) and Bayes (to classify the features in the fitness function) algorithms, Liu et al. [16] detected 18 genes that can be used for survival and/or prognostic factor in two groups of patients (less than 30 months or higher than 70 months) with colorectal cancer. Also, for the identification of biomarkers, some researchers have used Hierarchical Clustering [17] instead of Genetic Algorithms.

In conclusion, it is clear that there is no study that uses a combination of genetic and clinical information (heterogeneous data) in the characterization of cancer patients. Also, it is important to note that the majority of the studies did not use incomplete data in their works, which directly influences the complexity of any work. Finally. and after a careful literature analysis, the authors did not find any work that uses an evolutionary approach to clustering cancer patients, which constitute a novelty of this work.

III. Methodology

As previously mentioned, this paper reposts an initial phase of a large project with the final goal of predicting overall survival in breast cancer patients in a real clinical environment. Encompassed in the project architecture (Figure 1), in this phase a patient personalization step is presented, which consists in the detection of the different patient groups in the previously defined and constructed clinical database. The different stages of the global architecture are explained below:

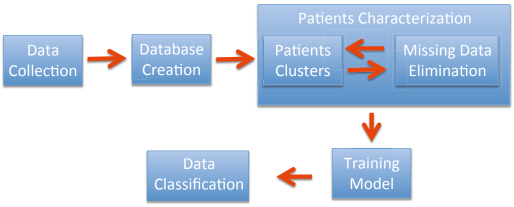

Fig. 1: Project Architecture.

1. **Data Collection**: The data was collected by a team composed by 4 medical doctors and includes information from 847 patient files with breast cancer from the same oncological center. Also, it is important to state that two other medical doctors performed a cross validation in the collected data in order to minimize the error introduced in this process. Each patient is characterized by 15 variables, including age, tumor site and topography, contralateral breast involvement, tumor stage (according to [18]), variables included in TNM classification (T: tumor size, N: nodes involved, M: metastasis), histological type, degree of differentiation, expression of hormonal receptors, expression of HER2 and type of treatment (including type of surgery, chemotherapy regimen, type of hormonotherapy, if applied).
2. **Database Creation**: After selecting and processing the patient files, a dataset was created to store all the data. Also, in this step, a team of two medical doctors performed the cross validation in the stored data.
3. **Patient Characterization – Patient Clusters**: The nature of the 15 variables stored in the database (linear and nominal) invalidates the use of Euclidean distance to measure the similarity of the data. In consequence, and as to characterize the patients, three algorithms were used – Hierarchical Clustering, k-medoids, and a Genetic Algorithm – using the distance measure proposed by Wilson for heterogeneous data [19]. This function (Equation 1) defines the distance between two values x and y of a given attribute a as:

$$d_a(x,y) = \begin{cases} 1 & x \text{ or } y \text{ unknown} \\ normalized\_vdm_a(x,y) & a \text{ is nominal} \\ normalized\_diff_a(x,y) & a \text{ is linear} \end{cases} \tag{1}$$

If some of the values are unknown the function will return a distance of 1. Otherwise, all data will be normalized and if the variables are nominal, Equation 2 was used,

$$normalized\_vdm_a(x,y) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2} \tag{2}$$

where:

- $N_{a,x}$ is the number of instances in the training set T that have value x for attribute a;

- $N_{a,x,c}$ is the number of instances in T that have value x for attribute a and output class c – in this case the output class was the overall survival;
- C is the number of output classes in the problem domain;

The other scenario relates to linear variables (Equation 3) where σa is the standard deviation of the numeric values of attribute a.

$$normalized\_diff_a(x,y) = \frac{|x-y|}{4\sigma_a} \quad (3)$$

In order to validate the results produces by the three algorithms, an evaluation function was created (Equation 4) aiming to maximize the distance between patient groups (intra-group distance) and minimize the distance between patients of the same group (inter-group distance).

$$F_{Eval} = \frac{Avg\,(Intra\text{-}Group\ Distance)}{Avg\,(Inter\text{-}Group\ Distance)} \quad (4)$$

The other three architecture steps (Patient Characterization – Missing Data Elimination, Training Model and Data Classification) were not treated in this study and therefore will not be described herein.

IV. Experimental Results

To allow for a comparison to be made between the used approaches, all of the three were tested varying the number of clusters between 1 and 200. For the genetic algorithm, 50 runs were executed for each configuration, considering a population of 1000 individuals and 25000 generations. The mutation operator was a change in cluster for a given patient and the crossover operator was a recombination of two chromosomes (population individuals). A typical run of our evolution algorithm is illustrated in Figure 2 where the x axis represents the number of generations and the y axis represents the evaluation function score (Equation 4). A more abrupt decrease can be seen in the initial generations with a estabilization occurring in later generations. The average evaluation score obtained by the evolutionary approach was 0.375, while the hierarchical approach achieved an average of 0.201 and the K-medoids obtained an average of 0.632 (Table 1). In order to compare the performance of the three algorithms, and given the small number of classifiers used [21], Tukey's HSD (Honestly Significant Difference) Test [20] with a significance level of 95% was used. This method compares each pair of algorithms (in this particular case 3 pairs) having as null hypothesis the equality of each pair. The results show that the comparison between k-medoids approach and each of the other two (Genetic and Hierarchical) presents a statistical difference with a $p-value < 0.0000001$. The other scenario (Genetic vs Hierarchical) also presents a statistical difference with a $p-value = 0.0061354$ (Table 2).

Table 1: Average Evaluation Function Score attending to Equation 4.

	Genetic	Hierarchical	k-medoids
Score	0.375	0.201	0.632

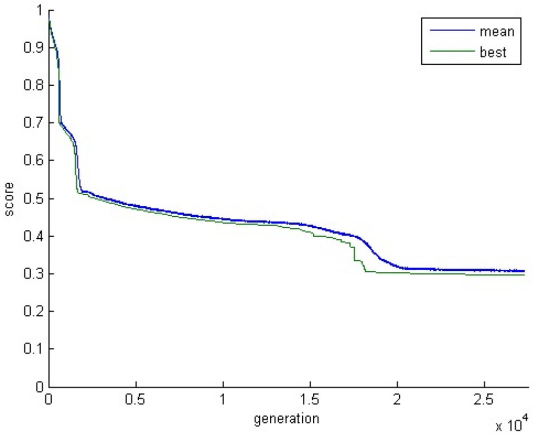

Fig. 2: Genetic Evolution.

Table 2: Tukey's HSD test results for the three pairs (G-H – Genetic/Hierachical, G-k-m – Genetic/K-medoids, H-k-m – Hierarchical/K-medoids.

	Tukey's HSD test value
G - H	0.0061354
G - k-m	<0.000001
H - k-m	<0.000001

The comparatively bad results achieve by K-medoids can be partially explained by the difficulty in obtaining a function to evaluate the distance between data points (patients) and in particular to determine the location of a new data point as the average location of a collection of data points. In what concerns to the performance of the evolutionary and hierarchical clustering, based on the achieved results, these can be in part explained by the fact that evolutionary approaches need additional computational resources and time to achieve the same level of result. In spite of the fact that at the beginning of the experimental setup 25000 generations were deemed enough to generate a good patient personalization, the obtained results, as can be seen in Figure 2, suggest that the use of a larger number of generations would bring forward better results.

V. Conclusions and Future Work

Encompassed in a more ambitious project, in this paper a comparison between three distinct algorithms to personalize breast cancer patients was performed. As far as the authors know, this was the first time that an evolutionary approach was used to cluster cancer patients. Regarding the obtained results, the three algorithms presented different performance, which can be due to many factors, as mentioned in the previous section. However, and regarding the evolutionary approach, extra time to perform additional experiments (which is usually seen as the main drawback in this type of approaches) would suggest an improvement of the results. Future work will focus on the remaining steps that were not described in section III – Methodology. The next step will be the elimination of the missing data from the database. This process will cover the choice of the algorithms and a validation step which includes monitoring the performance of those algorithms – this process will be accomplished by comparing the performance of the algorithms in a context where there is no missing data – from the literature analysis, this study was never performed. The other part of the validation process will consists in the comparison between the initial patient groups (performed in this paper) and the ones detected after the elimination of the missing values. This validation is represented in Figure 1 with the reciprocal relation between Patients Clusters and Missing Data Elimination. After that, classification algorithms will be used to train the model and to predict the overall survival of the breast cancer patients.

Acknowledgements

This work is partially supported by iCIS project (CENTRO-07-ST24-FEDER-002003) which is co-financed by QREN, in the scope of the Mais Centro Program and FEDER.

References

1. Nichols M., Townsend N., Luengo-Fernandez R., et al. European Cardiovascular Disease Statistics 2012 tech. rep.European Heart Network, Brussels 2012.
2. Siegel R., Naishadham D., Jemal A.. Cancer Statistics, 2013 *CA: A Cancer Journal for Clinicians.* 2013;63:11–30.
3. Slamon D, Eiermann W, Robert N, et al. Adjuvant Trastuzumab in HER2-Positive Breast Cancer *The new England J. of Medicine.* 2011:1273–1283.
4. Mesquita JM Bueno, Harten WH, Retel VP, et al. Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER) *Lancet Oncology.* 2007;8:1079–87.
5. Slodkowska EA, Ross JS. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients *J. of Expert Review of Molecular Diagnostics.* 2009;9:417–422.
6. Williams C, Brunskill S, Altman D, et al. Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy *J. of Health Technology Assessment.* 2006;10:1–204.
7. Cruz J A, Wishart D S. Applications of Machine Learning in Cancer Prediction and Prognosis *J. of Clinical Informatics.* 2007;2:59–77.
8. A. Endo H. Tanaka. Comparison of Seven Algorithms to Predict Breast Cancer Survival *Biomedical Soft Computing and Human Sciences.* 2008;13:11–16.
9. Wang K-M., Makond B., Wu W-L., Wang K-J, Lin Y.. Optimal data mining method for predicting breast cancer survivability *J. of Innovative Management,Information.* 2012;3:28–33.
10. Abreu P Henriques, Amaro H, Silva D Castro, et al. Overall Survival Prediction for Women Breast Cancer using Ensemble Methods and Incomplete Clinical Data in *IFMBE Proc, XIII Mediterranean Conference on Medical and Biological Engineering and Computing*:4 2013.
11. Fan G., Filipczak L., E.Chow . Symptom clusters in cancer patients: a review of the literature *Current Oncology.* 2007;14:173–179.
12. Husain A., Myers J., Selby D., Thomson B., Chow E.. Subgroups of Advanced Cancer Patients Clustered by Their Symptom Profiles: Quality-of-Life Outcomes *Journal of Palliative Medicine.* 2011;14:1246–1253.
13. Wang J., Liang X., Zhang Q., Fajardo L., Jiang H.. Automated breast cancer classification using near-infrared optical tomographic images *Journal of Biomedical Optics.* 2008;13.
14. Keskin F., Suhre A., Kose K., Ersahin T., Cetin A., Cetin-Atalay R.. Image Classification of Human Carcinoma Cells Using Complex Wavelet-Based Covariance Descriptors *PLoS ONE.* 2013;8.
15. Shah S., Kusiak A.. Cancer gene search with data-mining and genetic algorithms *Computers in Biology and Medicine.* 2007;37:251–261.
16. Liu Y., Aickelin U., Feyereisl J., Durrant L.. Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data *Knowledge Based Systems.* 2013;37:502–514.
17. Chen C-Y., Chou W-C., Tsay W., et al. Hierarchical cluster analysis of immunophenotype classify AML patients with NPM1 gene mutation into two groups with distinct prognosis *BMC Cancer.* 2013;13:1–9.
18. Edge S B, Byrd D R, Carducci M A, et al. , eds.*AJCC Cancer Staging Handbook.* Springer-Verlag New York Inc. 2009.
19. Wilson D., Martinez T.. Improved heterogeneous distance functions *J. of Artificial Intelligence Research.* 1997;6:1–34.
20. Zar J.. *Biostatistical Analysis.* Prentice Hall4th ed. 1999.
21. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets *J. of Machine Learning Research.* 2006;7:1–30.

Author: Pedro Henriques Abreu
Institute: Center for Informatic and Systems (CISUC)
Street: Pólo II, Pinhal de Marrocos
City: Coimbra
Country: Portugal
Email: pha@dei.uc.pt