Overall Survival Prediction for Women Breast Cancer using Ensemble Methods and Incomplete Clinical Data

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

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

Abstract— **Breast Cancer is the most common type of cancer in women worldwide. In spite of this fact, there are insufficient studies that, using data mining techniques, are capable of helping medical doctors in their daily practice.**

This paper presents a comparative study of three ensemble methods (TreeBagger, LPBoost and Subspace) using a clinical dataset with 25% missing values to predict the overall survival of women with breast cancer. To complete the absent values, the k-nearest neighbor (k-NN) algorithm was used with four distinct neighbor values, trying to determine the best one for this particular scenario. Tests were performed for each of the three ensemble methods and each k-NN configuration, and their performance compared using a Friedman test. Despite the complexity of this challenge, the produced results are promising and the best algorithm configuration (TreeBagger using 3 neighbors) presents a prediction accuracy of 73%.

Keywords— **Ensemble Methods, Overall Survival Prediction, Classification, Women Breast Cancer Dataset**

I. INTRODUCTION

Nowadays, cancer is one of the leading causes of death worldwide. According to Siegel [1], more than 1 million new cancer cases will be diagnosed and more than 580 thousand cancer deaths will occur in 2013 in the United States alone. Breast cancer is the most common cancer in women, and accounts for 29% of all cancer cases.

In the literature, some prognostic factors were described that influenced clinic decision. Patients with large tumors, not well differentiated, with no expression of hormonal receptors are expected to have worse prognosis and were treated more aggressively. These tumors are particularly prevalent in young patients (women less than 35 years old) and in spite of many developments in this area, this group is still a special one [2]. The research mark in the past two decades was the discovery of HER2, which showed that patients' treatment must be supported by a molecular understanding of breast tumors. This new marker was only detected in almost 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). Over the last 30 years, more than 3 million studies regarding cancer were conducted (values obtained using the ISI web of knowledge). However, in 2007, and according to Cruz and Wishart [7], less than 120 articles were related to cancer prediction/prognosis using soft-computing techniques. In this work, a survival model is presented based on 15 variables that are available in clinical practice. The challenge of this research is to understand if ensemble methods and k-NN algorithm can be used to create accurate predictors, in an oncological center using 847 patient files where 25% of the values are missing. The percentage of missing values found in the patient files reflects the reality of an oncological center that still uses physical patient files and constitutes by itself a good research challenge. The results are based in three different ensemble methods with different strategies in the classification process, and not only they show to be promising, but also open new perspectives for future works in the area.

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.

II. LITERATURE REVIEW

Over the years many studies have been developed in the area of cancer. Following the classification proposed by Cruz and Wishart [7], the cancer research area can be divided into cancer prediction and prognosis or cancer detection and diagnosis. As the work presented in this article is based on data collected from an oncological center, the second research area (cancer detection and diagnosis) will not be subject to analysis. In the area of cancer prediction and prognosis, many studies appeared over the last two decades. In spite of the fact that these studies are difficult to compare, mainly because they present different characteristics, such as number of cases to be examined or type of abnormalities, among others, we decided to divide the studies into classification of the tumor based in different types of clinical techniques (X-ray [8], microarray techniques [9]) or prediction, including cancer risk or susceptibility [10], cancer survivability [11] and cancer recurrence [12]. Having the main goal of this project in mind, in the breast cancer area there are still few works that used data mining techniques to predict patient survival [13]. In this work, the authors presented a comparison study that tried to predict patient survival using more than 200 thousand files and three different algorithms: Naive Bayes, Artifical Neural Networks and C4.5. However, this work presented some important drawbacks: authors eliminated incomplete data from the database, which substantially decreased the size of the original database; patient files included patients from different countries and some have more that 40 years, which means that many of the 16 variables used would already be outdated. At the end of the process, none of the algorithms proved to be better than the other. Similar issues are presented in the work presented by Endo et al. [14]. In this work, authors used the same database (provided by SEER - Surveillance Epidemiology and End Results [1]) and performed a comparison between Naive Bayes, Decision Trees (ID3 and J48) and a combination between Naive Bayes and Decision Trees. Also, the authors used only 10 variables in order to characterize the patient (most of then not clinical) and the range of the accuracy results revolves around 80%, which is far from ideal given that they eliminated noise from data at the beginning of the process. Finally, and following the same research line, Wang et al. [15] proposed a new method to predict breast cancer patients' survival using the SEER dataset. Doing a comparison with the other two analyzed studies, Wang improved the results to 90% of accuracy, but with the other detected issues still remaining.

In conclusion, and in spite the fact that some research studies addressed the problem of predicting breast cancer patient survival, none of the studies used only data from one oncology center and updated patient data; none of the studies used exclusively clinical variables; and none of the studies used incomplete data and ensemble methods in the prediction process, which constitutes the main contributions of this project.

[1] available at `http://www.seercancer.gov`

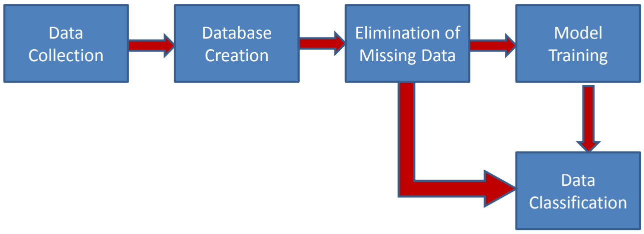

Fig. 1: Project Architecture

III. Methodology

The goal of this work is to predict the overall survival of women with breast cancer having as a base a dataset composed by more than 840 patient files. Generically, the goal is to identify which is the ensemble method that presents better performance classifying a breast cancer dataset with incomplete data. Having this goal in mind, 5 steps were defined (Figure 1):

1. **Data Collection:** The data was collected by a team composed by 4 medical doctors that collected information from 847 patient files with breast cancer over 2 months 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 in this process. Each patient was characterized by 15 variables, including age, tumor site and topography, contralateral breast involvement, tumor stage (according to [16]), 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 patient data. Also in this step, a team of two medical doctors performed the cross validation in the stored data.
3. **Elimination of Missing Data:** As often happens in clinical environments, some processes did not contain all patient information. This can be a result of many factors as explained in [17]. Analyzing more deeply the dataset produced in the previous step, we observe that 25% of the data was missing. To solve this problem, many strategies can be used, e.g omitting the instances with missing values, which is far from ideal, or using an algorithm in order to complete such data. In this project, the k-NN algorithm with four distinct values for neighbor (3, 5, 10 and 20) was used to complete the missing data by de-

tecting similarity between data. The choice of the algorithm is based in its implementation simplicity [18] and its good performance in such contexts [17]. At the end of this process, the neighbor value that minimizes classifier error will be detected. Finally, it is important to state that by the end of this step the dataset was split into two groups: the first group, composed by 240 randomly selected patients, will be used as new instances in the classification process; the second group (the remaining 607 patients) will be used in the next step (model training).

4. **Model Training:** To construct a classifier model, three ensemble methods were tested. Over the last decade, ensemble methods have proven themselves to be very effective and extremely versatile in a broad spectrum of problem domains and real-world applications [19]. For this project, three distinct methods from distinct data mining families were used: TreeBagger, LPBoost and Subspace.
5. **Data Classification:** After the construction of the ensemble models, the dataset with 240 patients (produced at the end of step 3) was used to analyze the classification performance of each ensemble method.

IV. Experimental Results

To produce the experimental setup of this work, the dataset produced in step 5 (explained in the previous section) was used. Due to the high percentage of missing data, the k-NN algorithm was used with four distinct values (3, 5, 10 and 20) for k, which produces not one but four distinct datasets to test. As 3 ensemble methods were used, that resulted in 12 distinct comparisons. To compare the 3 ensemble methods, the Friedman rank test was used. The averages of the results of each of the four configurations per algorithm were compared. The 240 patients in each of the four produced datasets were divided into 12 distinct groups (each group containing 20 randomly selected patients). The obtained ranks are shown in Table 1, where 'Number of NN' means number of nearest neighbors used and the number of the ranks varies between 1 (highest accuracy) and 12 (lowest one). Finally, in the case of a draw, average ranks are assigned [20]. Following the work presented by Demsar [20], and as $N > 10$ (number of split groups – 12) and $k > 5$ (number of classifiers used – 12) the proposed F_f value was calculated (7.34) and compared to the F distribution F(0.05) = 2.69. As a consequence, the null hypothesis of equivalence between the twelve predictors is rejected. Comparing the twelve configurations (four for each ensemble) for a 5% significance level using the Nemenyi test [20], it was possible to obtain CD = 4.810366. The CD is the critical value for the difference of mean ranks between the twelve predictors. It was proved that TreeBagger (NN–3) and LPBoost (NN–20) performed better than TreeBagger (NN–5) and Subspace (NN–5 and 10). Also, Subspace (NN–20) presented better performance comparing to TreeBagger (NN–5) and Subspace (NN–10). Finally, TreeBagger (NN–20) presented better performance than Subspace (NN–10).

Regarding to the classification performance and attending exclusively to the two algorithms that presented the highest mean in the Friedman Table (Table 1), TreeBagger (NN–3) presented 73% and LPBoost (NN–20) presented 70% of median concerning to hit rate in classification process, which constitutes a good and promising result, attending to the high percentage of missing data (25% of the values). Albeit previous results [14] [15] attained higher accuracy ratings, in these previous studies entries with missing values were eliminated, which obviously leads to better performance. However, it is important to stress that in the real world most of the patient files will be incomplete and therefore those studies do not reflect what can actually be achieved in practice, but rather what can be achieved in an idealistic scenario. This study reflects what can be achieved in the real world and presents a valid solution for overcoming the problem of missing data.

V. Conclusions and Future Work

In this research work, an overall survival prediction approach for the most common cancer pathology in women (breast cancer) was presented. Based on 847 patients files collected at the same oncological center, the performance of three ensemble methods was compared. The results showed that, even with a high percentage of missing values (around 25%), it is possible to obtain good results in the prediction of overall survival.

Further developments in this research project shall focus in several distinct areas: increase the number of patients or expand the study to predict the disease free survival. The first identified direction will have huge similarity with the presented project concerning to the collection of the data and it will be very interesting if the new study focuses exclusively in a group of patients, e.g. younger ones rather than englobing patients with a wide range of ages, allowing the identification of features that most influence survival in those groups. The second future direction consists in expanding this study to predict the disease free survival. Nowadays, and fortunately in some cases, the survival of a cancer patient is very good; however, some tumors recur over time. Because of that, it is important to know what the disease free survival of a patient would be. At the end of this study, a new study can emerge in the area of optimization problems. Combining overall survival and the free survival, the goal is to find the treatment

Table 1: Ranks of the Friedman test for the three ensemble method and k-NN configuration (Number of NN) for each of the 12 groups of patients. The last column presents the mean ranking accross the 12 groups

		Groups												
Ensemble Method	Number of NN	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**MEAN**
TreeBagger	3	2.5	1	3	5.5	1.5	6.5	5.5	3	2	5.5	1	3.5	3.375
	5	12	12	11.5	12	10.5	4.5	7.5	11.5	8	2.5	11	11	9.500
	10	2.5	5.5	1	10.5	10.5	1	12	10	2	2.5	5	12	6.208
	20	5.5	4	3	2.5	5.5	12	1	2	4.5	5.5	11	3,5	5
LPBoost	3	2.5	5.5	6	8.5	1.5	6.5	9	4.5	4.5	9	6.5	10	6.167
	5	7.5	2.5	6	5.5	9	9.5	10	8	8	9	11	6	7.667
	10	7.5	11	9	1	4	8	3	7	2	5.5	2	8	5.667
	20	2.5	2.5	3	2.5	7	3	2	6	6	5.5	3.5	5	4.042
Subspace	3	9	8	6	7	5.5	4.5	7.5	4.5	11	9	6.5	1.5	6.667
	5	11	10	10	8.5	8	9.5	5.5	9	8	11	8	8	8.875
	10	10	9	11.5	10.5	12	11	11	11.5	12	12	9	8	10.625
	20	5.5	7	8	4	3	2	4	1	10	1	3.5	1.5	4.208

that optimizes both previously defined target functions, supporting the clinician in his treatment decision.

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. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013 *A Cancer J. for Clinicians.* 2013;63:11–30.
2. Banz-Jansen C, Heinrichs A, Hedderich M, et al. Are there changes in characteristics and therapy of young patients with early-onset breast cancer in Germany over the last decade? *Archives of Gynecology and Obstetrics.* 2013:1-5.
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.* 2017;2:59–77.
8. Vasantha M, Bharathi V, Dhamodharan S. Medical Image Feature, Extraction, Selection And Classification *International J. of Eng. Science and Technology.* 2010;2:2071–2076.
9. X Ruan J Wang, Li H, Li Xiaoming. A Method for Cancer Classification Using Ensemble Neural Networks with Gene Expression Profile in *Conference on Bioinformatics and Biomedical Eng.*:342–346 2010.
10. Dumitrescu R, Cotaria I. Understanding breast cancer risk where do we stand in 2005? *J. of Cellular and Molecular Medicine.* 2005;9:208–211.
11. Futschik M E, Kasabov N, Reeve A, Sullivan M. Prediction of clinical behavior and treatment for cancers *J. of Applied Bioinformatics.* 2003;2:53–58.
12. Fan Q, Zhu C-J, Yin L. Predicting breast cancer recurrence using data mining techniques in *International Conference on Bioinformatics and Biomedical Technology (ICBBT)*:310–311 2010.
13. Sarvestani A S, Shiraz I, Safavi A A, Parandeh N M, Salehi M. Predicting breast cancer survivability using data mining techniques in *2nd International Conference on Software Technology and Engineering (ICSTE)*:227–231 2010.
14. Endo A, Shibata T, Tanaka H. Comparison of Seven Algorithms to Predict Breast Cancer Survival *Biomedical Soft Computing and Human Sciences.* 2008;13:11–16.
15. Wang K-M, Makond B, Wu W-L, Wang K-J, Lin Y S. Optimal data mining method for predicting breast cancer survivability *International J. of Innovative Management,Information.* 2012;3:28–33.
16. Edge S B, Byrd D R, Carducci M A, et al. , eds.*AJCC Cancer Staging Handbook*. Springer-Verlag New York Inc. 2009.
17. Twala B, Cartwright M, Shepperd M. Comparison of various methods for handling incomplete data in software engineering databases in *International Symp. on Empirical Software Eng., 2005*:10 pp 2005.
18. Jain A. Data clustering: 50 years beyond K-means *Pattern Recognition Letters.* 2010;31:651–666.
19. Zhang C, Ma Y. , eds.*Ensemble Machine Learning*. Springer-Verlag New York Inc. 2012.
20. 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