Using Supervised Machine Learning for Breast Cancer Diagnosis

Michael Tutterrow
Computer and Mathematical Sciences
Lewis University
michaelatutterrow@lewisu.edu

Ricardo Morales Ballesteros
Computer and Mathematical Sciences
Lewis University
ricardomoralesbal@lewis.edu

***Abstract* — Breast cancer has become the primary cause of death in women from developed countries, increasing significantly during the last few decades, becoming the second most common cause of death in women worldwide.**

Early diagnosis of breast cancer is essential to save lives of patients. Usually, medical datasets include a large variety of data that can lead to confusion during diagnosis. It requires elimination of inappropriate and repeated data from the dataset before final diagnosis. This can be done using any of the feature selection algorithms available in data mining. Feature selection is considered as a vital step to increase the classification accuracy.

For this classification task, the WBCD (Wisconsin Breast Cancer Diagnosis) dataset is employed [1]. This dataset is widely utilized for this kind of application because it has a large number of instances (699), is virtually noise-free and has just a few missing values [3].

This report documents the analysis of a dataset used for breast cancer diagnosis. It is important to highlight that much of the work implemented before performing the tests was dedicated to pre-processing of the data in order to optimize the classifier. Issues addressed include dealing with missing data and avoiding overfitting of the classifier.

The results are presented in tabular format, which contains the accuracy of the classifier as well as the key performance metrics of each one of the algorithms. All the tests were conducted using the software Weka, an open-source collection of machine learning techniques capable of performing pre-processing, classification, regression, clustering and association rules.

Keywords—supervised machine learning, data mining, analytics, Weka, breast cancer diagnosis

I. OBJECTIVE

Breast cancer is the most common cancer among women in the United States, with over 266,000 new cases expected for the year 2018. The most common symptoms of breast cancer include a lump in the breast or armpit, a change in breast size or shape, fluid coming from the nipple, and red peeling skin. Risk factors for breast cancer include genetics, obesity, alcohol consumption, hormone therapy, and age [2].

Breast cancer is typically detected either during a screening examination, before symptoms have developed, or after a woman notices a lump. Most masses seen on a mammogram and most breast lumps turn out to be benign (not cancerous), do not grow uncontrollably or spread, and are not life-threatening. When cancer is suspected, microscopic analysis of breast tissue is necessary for a diagnosis and to determine the extent of spread (stage) and characterize the type of the disease. The tissue for microscopic analysis can be obtained from a needle biopsy (fine-needle or wider core needle) or surgical incision. Selection of the type of biopsy is based on multiple factors, including the size and location of the mass, as well as patient factors and preferences and resources.

According to the American Cancer Society, relative survival rates for women diagnosed with breast cancer are [4]:

- 91% at 5 years after diagnosis.
- 86% after 10 years
- 80% after 15 years

The above data is a clear indicator of the impact on early detection of breast cancer; detected in its early stages, there is a 30% chance that the cancer can be treated effectively, but the late detection of advanced-stage tumors makes the treatment more difficult [3].

Currently, the most used techniques to detect breast cancer in early stages are: mammography (63%to 97% correctness), FNA (Fine Needle Aspiration) with visual interpretation (65% to 98% correctness) and surgical biopsy (approximately 100% correctness). Therefore, despite the wide variation in correctness for the mammography and FNA with visual interpretation, both are the quickest and most accessible mechanisms for detection, considering that the surgical biopsy, although reliable, is invasive and costly [2].

Several papers have been published during the last 20 years trying to achieve the best performance for the computational interpretation of digitalized imaging of FNA and mammography, as part of the screening examination and testing for early breast cancer detection.

In this report, we discuss and evaluate the use of two supervised learning algorithms. We use the "Logistic Regression" implemented in Weka with the *SimpleLogic*

function as well as a “Decision Tree” using the C4.5 algorithm, implemented by the function *J48* [5,6].

A. *Dataset*

The dataset used in this paper is publicly available [1] and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin,USA. It was donated by Olvi Mangasarian on July 15th,1992.

Samples arrived periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself [1]:

Group 1: 367 instances (January 1989)

Group 2: 70 instances (October 1989)

Group 3: 31 instances (February 1990)

Group 4: 17 instances (April 1990)

Group 5: 48 instances (August 1990)

Group 6: 49 instances (Updated January 1991)

Group 7: 31 instances (June 1991)

Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated datbase on 15 July 1992)

To collect and consolidate the data here, Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt [3].

The program uses a curve-fitting algorithm, as shown in Fig. 1, to compute ten features from each one of the cells in the sample.

Fig. 1. A magnified image of a malignant breast fine needle aspirate.

Each feature is evaluated on a scale of 1 to 10, with 1 being the closest to benign and 10 the closest to malignant, including a class attribute which is defined to identify a benign (2) or malignant (4) mass, plus a sample code number that is simply an identification for each record.

The statistical analysis done during the consolidation of the data, showed that the following 9 features differ significantly between benign and malignant samples:

- uniformity of cell shape
- uniformity of cell size
- clump thickness
- bare nuclei
- cell size
- normal nucleoli
- clump cohesiveness
- nuclear chromatin
- mitoses

The data can be considered ‘noise-free‘ and has 16 missing values, which are the Bare Nuclei for 16 different instances. Table 1 is a summary of the current state of the dataset used in this report [3].

TABLE 1. DATASET SUMMARY

Features	Uniformity of cell shape	Numeric	1-10
	Uniformity of cell size	Numeric	1-10
	Clump thickness	Numeric	1-10
	Bare nuclei	Numeric	1-10
	Cell size	Numeric	1-10
	Normal nucleoli	Numeric	1-10
	Clump cohesiveness	Numeric	1-10
	Nuclear chromatin	Numeric	1-10
	Mitoses	Numeric	1-10
	Class	Nominal	Benign, Malignant
Class Distribution		Benign: 458 (65.5%)	
		Malignant: 241 (34.5%)	
Number of Missing Values		16	
Number of Instances		699	

II. DATA MINING PROCESS

The initial step to be taken here is the pre-processing of the data, using the tools available in Weka [5,6]. Considering the dataset in use, the pre-processing is focused on managing the missing attributes.

To manage the 16 missing values, two methods are proposed: the first one is to use the filter *ReplaceMissingValues*. This filter will replace all missing values for attributes in the dataset with the means from the training data [5,6]. The second option is to remove all the instances with missing values, and the new dataset will have 683 instances.

Considering the nature of the missing attributes (all are “bare nuclei size”) the first impression is that using the filter *ReplaceMissingValues* is not a good option, because the size of an individual cell is not related to the mean size of the other cells.

With the above consideration, the option taken was the removal of the records with missing values, leaving our new dataset with 638 instances, as described above.

A second condition observed was the distribution of the data collected. In our scenario, with the present dataset, the inspection shows that the number of observations belonging to one class is significantly lower than those belonging to the other classes. Using Weka, we were able to visually observe this imbalance condition, as presented on the Fig. 2.

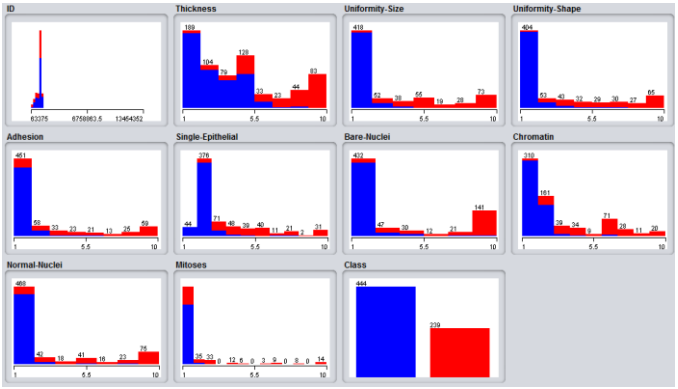

Fig. 2. WBCD Dataset – Evaluation of attributes shows the imbalance scenario.

When studying problems with imbalanced data, it is crucial to adjust either the classifier or the training set balance, or even both, to avoid the creation of an inaccurate classifier.

In this report, the problem with the imbalanced data is solved by choosing machine learning methods that are insensitive to this kind of issue. For this purpose, we have chosen the *SimpleLogistic* algorithm as part of the Logic Regression function and the *J48*, which is a reimplementation of C4.5.

A. *Logistic Regression*

A Logistic Regression algorithm helps to find a function between dependent variable (which has to be categorical variable) and independent variable(s) (which can be continuous or categorical variable). It is a linear classifier which means that decision boundary is linear. It is used to solve classification problems.

There are two logistic regression functions available in Weka. *SimpleLogistic* uses LogitBoost with simple regression functions as base learners. Cross validation is used to determine how many iterations to perform, and also supports automatic attribute selection. *Logistic* uses a ridge estimator to avoid overfitting by penalizing large coefficients [6]. *SimpleLogistic* results in slightly higher predictive accuracy for this dataset (~1%), and is documented here.

Settings: Using the default setting of *useCrossValidation* to determine the number of LogitBoost iterations provides the highest accuracy. Settings for *errorOnProbabilities* and *useAIC* were also tested and found to be less accurate, or no different.

Training/Testing Split: The dataset is divided into 60/40 percentage splits for training/testing of the regression model.

B. *Decision Tree*

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

The Weka implementation for a decision tree using the C4.5 algorithm is the *J48* classifier in Weka [6]. This algorithm is used to model a decision tree for this dataset.

Settings: Default settings include the unpruned option set to False. This setting proved beneficial, as building an unpruned tree generated 50% more levels, with only a marginal increase (~1%) in accuracy.

Training/Testing Split: The dataset is divided into 60/40 percentage splits for training/testing of the regression model.

III. RESULTS

A. *Logistic Regression*

Classifier Model: Equations (1) and (2) show the results of the logistic regression model, for classification of classes 2 and 4 (benign and malignant), respectively.

(1)

```
4.8  +
[Thickness] * -0.26 +
[Uniformity-Size] * -0.16 +
[Uniformity-Shape] * -0.04 +
[Adhesion] * -0.14 +
[Single-Epithelial] * -0.05 +
[Bare-Nuclei] * -0.22 +
[Chromatin] * -0.16 +
[Normal-Nuclei] * -0.1 +
[Mitoses] * -0.21
```

(2)

```
-4.8 +
[Thickness] * 0.26 +
[Uniformity-Size] * 0.16 +
[Uniformity-Shape] * 0.04 +
[Adhesion] * 0.14 +
[Single-Epithelial] * 0.05 +
[Bare-Nuclei] * 0.22 +
[Chromatin] * 0.16 +
[Normal-Nuclei] * 0.1  +
[Mitoses] * 0.21
```

Confusion Matrix: Table 2 shows the confusion matrix for the 273 instances of the testing dataset.

TABLE 2. LOGISTIC REGRESSION CONFUSION MATRIX

Actual:	Predicted: Class = 2 (Benign)	Class = 4 (Malignant)
Class = 2	167	4
Class = 4	6	96

Performance Measures: Summaries of performance measures and accuracy for the testing dataset are shown in Tables 3 and 4.

TABLE 3. PERFORMANCE MEASURES

Performance Measures		
Total Number of Instances		273
Correctly Classified Instances (Accuracy)	(263)	96.34%
Incorrectly Classified Instances	(10)	3.66%
Kappa statistic		0.9214
Mean absolute error		0.0515
Root mean squared error		0.1718
Relative absolute error		11.24%
Root relative squared error		35.39%

TABLE 4. ACCURACY BY CLASS

Accuracy By Class	Precision	Recall	F-Measure	Class
	0.965	0.977	0.971	2
	0.96	0.941	0.95	4
Weighted Avg.	0.963	0.963	0.963	

B. *Decision Tree*

Decision Tree: The decision tree modeled by Weka is shown in Fig. 3.

Decision Tree Rules: Several rules can be extracted from each path in the decision tree, to create actionable knowledge about diagnosing a tumor. An example of a rule is shown in (3).

(3)

```
IF (Uniformity-Size <= 2) AND (Bare-Nuclei > 3)
AND (Thickness <= 3) THEN (Class = 2: Benign)
```

Confusion Matrix: Table 5 shows the confusion matrix for the 273 instances of the testing dataset.

TABLE 5. DECISION TREE CONFUSION MATRIX

Actual:	Predicted: Class = 2 (Benign)	Class = 4 (Malignant)
Class = 2	168	3
Class = 4	14	88

Performance Measures: Summaries of performance measures and accuracy for the testing dataset are shown in Tables 6 and 7.

TABLE 6. PERFORMANCE MEASURES

Performance Measures		
Total Number of Instances		273
Correctly Classified Instances (Accuracy)	(256)	93.77%
Incorrectly Classified Instances	(17)	6.23%
Kappa statistic		0.864
Mean absolute error		0.0739
Root mean squared error		0.2436
Relative absolute error		16.12%
Root relative squared error		50.20%

TABLE 7. ACCURACY BY CLASS

Accuracy By Class	Precision	Recall	F-Measure	Class
	0.923	0.982	0.952	2
	0.967	0.863	0.912	4
Weighted Avg.	0.94	0.938	0.937	

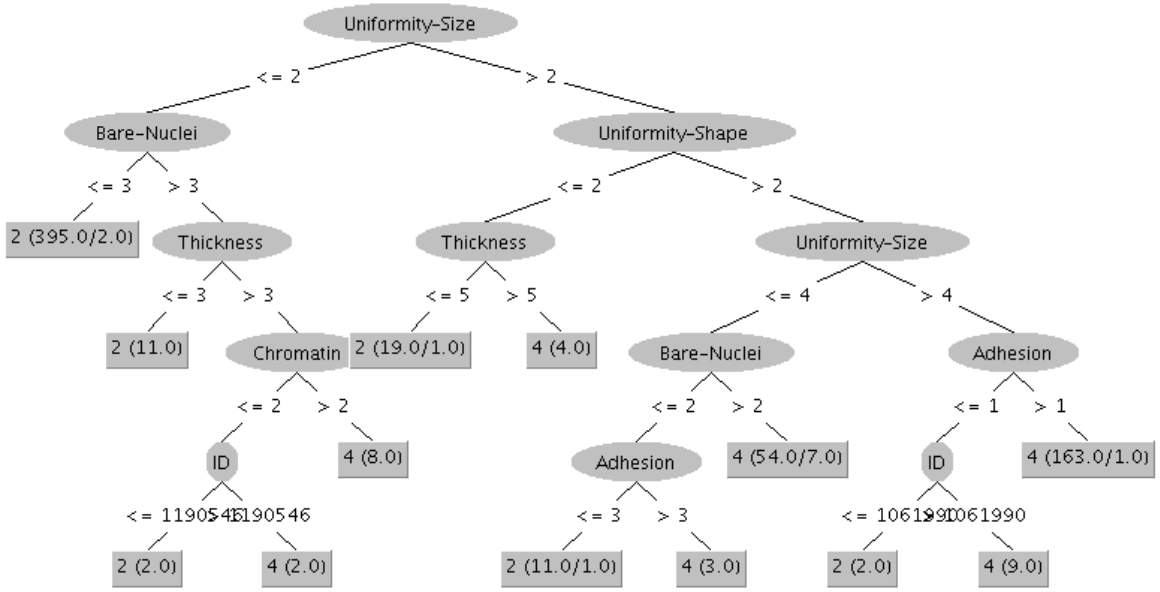

Fig. 3. C4.5 Decision Tree

IV. CONCLUSIONS

In this report, we analyzed and tested the use of two distinct machine learning techniques for breast cancer diagnosis. The first algorithm, Logistic Regression, demonstrated a good performance when dealing with imbalanced data (96.34% of accuracy). The other algorithm, a C4.5 Decision Tree, resulted in a less accurate classifier, with a higher rate of false-negatives when compared to the first one (93.77% of accuracy). Both classifiers have accuracies at the high end of the ranges for conventional methods previously mentioned, and present diagnostic processes that are easily implemented by health care practitioners.

Data mining can be a highly effective and practical approach towards breast cancer diagnosis. Early detection and diagnosis depends on a pathology report detailing specific attributes of the mass in question, with the appropriate application of machine learning methods to an existing dataset. This approach is shown to be one of the most accurate and effective diagnostic methods available.

REFERENCES

[1] William H. Wolberg, W. Nick Street, Olvi L. Mangasarian. "Breast Cancer Wisconsin (Diagnostic) Data Set". University of California, Irvine (UCI), Information and Computer Science, Machine Learning Repository. [Online].

Available:http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[2] American Cancer Society: American Cancer Society Recommendations for the Early Detection of Breast Cancer, 2017. [Online].

Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html#references

[3] Borges, Lucas Rodrigues, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection", Proceedings of XI Workshop de Visão Computacional, 2015, [Online].

Available:http://www.lbd.dcc.ufmg.br/colecoes/wvc/2015/001.pdf

[4] American Cancer Society. Breast Cancer Facts & Figures 2017-2018. Atlanta: American Cancer Society, Inc. 2017. [Online].

Available:https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.pdf

[5] University of Waikato, "Weka 3: Data Mining Software in Java,", 2018 [Online].

Available: https://www.cs.waikato.ac.nz/~ml/weka/downloading.html

[6] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

APPENDIX: TEAM MEMBER CONTRIBUTION

Both authors investigated several datasets, with preliminary analysis performed in Weka, Python, and Excel. M.T. performed the regression and decision tree analysis for the [1] dataset and wrote an initial report draft to summarize the findings. R.M.B. performed additional supporting research, attributes analysis, and wrote the report sections for the objective and data mining process. Both authors collaborated on the final draft of the report.

ACKNOWLEDGMENT

We would like to thank all the precedent contributions and work that led to the creation of the WBCD dataset, allowing us to develop the work seen on this report:

- O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
- William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
- O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
- K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).