Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection

Conference Paper · October 2015

CITATIONS
19

READS
20,962

1 author:

Lucas Borges
University of São Paulo
43 PUBLICATIONS **180** CITATIONS

SEE PROFILE

Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection

Borges, Lucas Rodrigues
Union College
P.O. Box 1916
Schenectady, New York
rodriguj@garnet.union.edu

ABSTRACT

Two machine learning techniques are compared in this paper. These methods are used to create two classifiers that must discriminate benign from malignant breast lumps. To create the classifier, the WBCD (Wisconsin Breast Cancer Diagnosis) dataset is employed. 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. Before performing the tests, a large fraction of this work will be dedicated for pre-processing the data in order to optimize the classifier.

The first part of this work is to present the datase, what it contains, when and how it was created, if it is noisy, if it has missing values. This section is important to understand what are the issues that will need to be processed while preparing the data to create the classifier.

The next step is to propose methods and algorithms to optimize the training set. How to deal with missing values? How to avoid overfitting the classifier? All these questions are discussed and different solutions are proposed.

The results are presented in tables, which contains the accuracy of the classifier, the rate of false-negatives and the rate of false-positives[1]. All the tests were conducted using the software Weka 3.6, an open-source collection of machine learning techiniques capable of performing pre-processing, classification, regression, clustering and association rules.

The best accuracy in this paper was achieved by the Bayesian Networks algorithm, wich had, in its best configuration, 97.80% of accuracy. The second algorithm tested was the J48, which had 96.05% of accuracy.

[1]Throughout this paper, the expression "False-Negative" is used to name the instances that were classified as Benign but in reality are malignant, and "False-Positive" is for the instances misclassified as Malignant

Keywords

Breast Cancer, Detection, Weka, FNA, Wisconsin, Dataset, Machine Learning, Bayesian Networks, J48.

1. INTRODUCTION

Breast cancer is the most common cancer among women and one of the major causes of death among women worldwide [1]. Every year approximately 124 out of 100,000 women are diagnosed with breast cancer, and the estimation is that 23 out of the 124 women will die of this disease [2].

When 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 [3,4]. Currently, the most used techniques to detect breast cancer in early stages are: mammography (63% to 97% correctness [5]), FNA (Fine Needle Aspiration) with visual interpretation (65% to 98% correctness [6]) and surgical biopsy (approximately 100% correctness). Therefore, mammography and FNA with visual interpretation correctness varies widely, and the surgical biopsy, although reliable, is invasive and costly.

This paper discuss a diagnosis technique that uses the FNA (Fine Needle Aspiration) with computational interpretation via machine learning and aims to create a classifier that provides a high level of accuracy, with a low rate of false-negatives.

Several papers were published during the last 20 years trying to achieve the best performance for the computacional interpretation of FNA samples[7], and in this paper two well-known machine learning techniques are tested: Bayesian Networks and J48.

Building a classifier using machine learning can be a difficult task if the dataset used is not on its best format or if it is not being correctly interpreted. Therefore, a considerable portion of this work will be spent preparing and comprehending the dataset in order to avoid problems such as overfitting. To prepare the dataset, the tab "Preprocess" implemented on Weka will be explored to find appropriate filters and prepare the training set before it can generate the classifier.

2. DATASET

The dataset used in this paper is publically available[8] 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 [9].

To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses[10] and an easy-to-use graphical computer program called Xcyt[11], which is capable of perform the analysis of cytological features based on a digital scan.

Table 1: Gruops of instances

Group 1	369 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	January 1991
Group 7	31 instances	June 1991
Group 8	86 instances	November 1991
Total: 701		

The program uses a curve-fitting algorithm, as shown in Figure 1, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

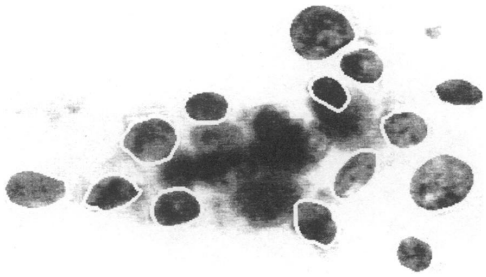

Figure 1: A magnified image of a malignant breast fine needle aspirate. Visible cell nuclei are outlined by a curve-fitting program. The Xcyt system also compares various features for each nucleus.

Each feature is evaluated on a scale of 1 to 10, with 1 being the closest to benign and 10 the closest to malignant. Statistical analysis[12] showed that the following nine characteristics 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 and mitoses[2].

The samples were taken periodically as Dr. Wolberg reported his clinical cases; therefore the data is presented as chronological groups that reflect the period they were created. Table 1 shows the number of instances added each month since the dataset started being built (Janurary 1989) until the last instance created (November 1991).

Before being publically available the dataset had 701 points, but on January of 1989, after being revised, 2 instances from group 1 were considered inconsistent and were removed from the dataset. Two more revisions occurred before the actual state of the dataset, both of them aimed to substitute values from zero to one, so the value range of the features is 1-10.

The data can be considered 'noise-free'[13] and has 16 missing values, which are the Bare Nuclei for 16 different instances, from group 1 to 6. Table 2 is a summary of the current state of the dataset used in this paper.

3. EXPERIMENTS

The first step of the experiment is pre-processing the data, using the tools available in Weka 3.6. Considering the dataset adopted, the pre-processing will focus on manage the missing attributes, the unbalanced data and the number of attributes used to train the classifier.

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. 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 of them are bare nuclei size) the first impression is that replacing them with the mean value from the training set is not a good option, because the size of an individual cell is not related to the mean size of the other cells. The next section will show that this idea is reflected into the classifier's performance.

A key issue in machine learning is to avoid "overtraining" the classifier, that is, memorizing details of the training data at the expense of good generalization to unseen data. In our dataset, a good generalization is achieved by reducing the number of input features. To decide which attributes are more relevant for the classification, the filter "InfoGainAttributeEval" was applied. This filter evaluates the worth of an attribute by measuring the information gain with respect to the class, and then it ranks the attributes by their individual evaluations. In order to find the best number of attributes, the classifiers are trained using different combinations of attributes, and the accuracy of each one is compared.

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. A common practice for dealing with imbalanced data sets is to rebalance them artificially, which is called "up-sampling" (replicating features from the minority) and "down-sampling" (removing cases from the majority). There are plenty of studies demonstrating that this kind of technique does not have a great effect on the predictive performance of learned classifiers [14].

In this paper, the problem with the imbalanced data is

[2]The dataset publically available provides only those nine features

Table 2: Sumary of the dataset

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	

solved by choosing machine learning methods that are insensitive to this kind of issue. The first is a Bayesian classifier called Bayesian Networks, which is commonly tolerant to imbalanced dataset [15]. To use this classifier, the dataset must not have any missing value and the attributes must be nominal [16]. Thus, before training the classifier the attributes are going to be discretized using the filter implemented in Weka.

The second learning algorithm is the J48, which is a reimplementation of C4.5. This algorithm has better performance dealing with imbalanced data if some of its attributes are configured correctly [17].

4. RESULTS

4.1 Bayesian Networks

To correctly analyze the results, is important to should keep in mind that for this application of machine learning, having an accurate classifier is as important as having a low rate of false-negative when classifying a malignant lump, because each instance miss classified as a benign lump can delay the correct diagnosis and turn the treatment even more difficult.

The first set of tests was made using the Bayesian Networks Algorithm, and the first stage was discretizing the attributes to improve the performance of the algorithm. To compare the impact of the discretization, the algorithm was tested with its original values and filtered with and without the use of equal frequency. The results are showed at Table X.

Table 3: Discretizing the dataset (BN)

	Accuracy	False-Negative	False-Positive
Original	97.14%	4 (0.57%)	16 (2.29%)
Discretized	97.28%	4 (0.57%)	16 (2.16%)
EqualFreq.	97.42%	3 (0.43%)	16 (2.16%)

Analyzing Table 3 is evident that both the best accuracy and the lowest false negative rate are obtained by applying the discretization filter with the equal frequency mode. All the subsequent tests are performed using the discretized dataset with equal frequency.

The next step is testing the two proposed methods for dealing with the missing attributes. Three different performances are compared: with no special treatment for the missing values, replacing the missing values for attributes in the dataset with the means from the training data or simply removing the instances that contain missing attributes.

Table 4: Manage the missing attributes (BN)

	Accuracy	False-Negative	False-Positive
No Filterl	97.42%	3 (0.43%)	15 (2.16%)
Replaced	97.42%	3 (0.43%)	15 (2.16%)
Removed	97.51%	4 (0.59%)	13 (1.90%)

Table 4 shows the results obtained by using 10 fold cross-validation. The best accuracy was achieved by removing the instances that contains missing attributes, but this method also had a higher rate of false negatives. Therefore, using just the results obtained on Table 4 it is not possible to confirm which method has the best performance, and the next steps will be conducted using both replacing and removing methods.

Table 5: Information Gain from all attributes

Rank	Attribute	Information Gain
1	Uniformity of Cell Size	0.701
2	Uniformity of Cell Shape	0.677
3	Bare Nucleoli	0.598
4	Bland Chromatin	0.555
5	Single Epithelial Cell Size	0.534
6	Normal Nucleoli	0.487
7	Clump Thickness	0.464
8	Marginal Adhesion	0.464
9	Mitosis	0.212

For the next experiment, the function “Select Attributes” is used to evaluate the worth of an attribute by measuring the information gain with respect to the class, and then it ranks the attributes by their individual evaluations. After generating the rank, presented at Table 5, each attribute is removed in order to achieve the best correctness and lowest false-negative rate.

Table 6 presents the results from the process of removing

Table 6: Removing Attributes (BN)

	Discretized (EqualFreq) + Replaced			Discretized (EqualFreq) + Removed		
Attributes #	Accuracy	False-Negative	False-Positive	Accuracy	False-Negative	False-Positive
9 + Class	97.42%	3(0.43%)	15(2.15%)	97.51%	4(0.59%)	13(1.90%)
8 + Class	97.42%	3(0.43%)	15(2.15%)	97.80%	2(0.29%)	13(1.90%)
7 + Class	97.42%	3(0.43%)	15(2.15%)	97.36%	5(0.73%)	13(1.90%)
6 + Class	97.42%	3(0.43%)	15(2.15%)	97.66%	3(0.44%)	13(1.90%)
5 + Class	96.99%	6(0.86%)	15(2.15%)	97.22%	6(0.88%)	13(1.90%)

attributes. The best performance was obtained using 8 attributes plus the class, with the missing values removed from the dataset. The removed attribute is "Mitosis", which is ranked at the last position.

4.2 J48

In order to find the best classifier, the same tests performed for the Bayesian Networks Algorithm are going to be repeated for J48, and like before the first test will compare the performance of the classifier when the dataset is discretized.

Table 7: Discretizing the dataset (J48)

	Accuracy	False-Negative	False-Positive
Original	94.56%	18(2.57%)	20(2.86%)
Discretized	94.42%	18(2.57%)	21(3.00%)
EqualFreq.	93.56%	19(2.72%)	26(3.72%)

From Table 7 is possible to infer that the data before being and after filtered results in similar performance of the classifier. For the next tests both possibilities are going to be considered, and after the next pre-processing steps it will be possible to recognize which option is the best.

The next step of the pre-processing is to manage the missing values. The options are, again, either to replace the missing attributes with the mean value calculated from the training data or simply remove incomplete instances from the training set.

Table 8: Manage the missing attributes (J48)

Original	Accuracy	False-Negative	False-Positive
No Filter	94.56%	18(2.57%)	20(2.86%)
Replace	95.14%	14(2.00%)	20(2.86%)
Removed	96.05%	11(1.61%)	16(2.34%)
Discretized	Accuracy	False-Negative	False-Positive
No Filter	94.42%	18(2.57%)	21(3.00%)
Replace	94.42%	18(2.57%)	21(3.00%)
Removed	93.41%	23(3.37%)	22(3.22%)

Analyzing Table 8 is possible to affirm that the dataset that had its instances with missing values removed and that was not discretized can generate a better classifier.

The last but not less important test is to use the function âĂIJSelect AttributesâĂİ to generate the attributes rank. The rank obtained for this configuration of the dataset is the same as presented in Table 5. The result of removing attributes is shown at Table 9 and the conclusion that can be drawn from the performance is that for this algorithm almost all the attributes have the same impact at the classifier's performance. It is interesting to note that the rate of false-negative for the classifier trained with 9 attributes plus the class is the same as the classifier trained with just 3 attributes plus the class.

Table 9: Removing Attributes (J48)

Attributes #	Accuracy	False-Negative	False-Positive
9 + Class	96.05%	11(1.61%)	16(2.34%)
8 + Class	95.75%	12(1.76%)	17(2.49%)
7 + Class	95.90%	12(1.76%)	16(2.34%)
6 + Class	95.17%	16(2.34%)	17(2.49%)
5 + Class	95.17%	14(2.05%)	19(2.78%)
4 + Class	95.17%	13(1.90%)	20(2.93%)
3 + Class	95.61%	11(1.61%)	19(2.78%)

5. CONCLUSION

In this paper we investigated the use of two distinct machine learning techiniques for breast cancer diagnosis. The first algorithm, Bayesian Networks, demonstrated a good performance when dealing with imbalanced data (97.80% of accuracy), but it is important that, before running the algorithm the dataset must be pre-processed, because it does not deal with missing values, and it has a better performance when learning from a dataset with discretized nominal values.

The other algorithm, J48, resulted in a less accurate classifier, with a higher rate of false-negatives when compared to the frist one (96.05% of accuracy). This behavior was expected, once this is a tree algorithm, and tree algorithms has worse accuracy when dealing with imbalanced datasets if compared to Bayers algorithms.

Table 10: Papers using the WBCD.

Algorithm	Accuracy	Year	Reference
Back-Propagation	94.90%	1992	[18]
MSM-T	97.00%	1993	[10]
MSM-T + Pre-Proc.	97.50%	1995	[11]
Fuzzy-Genetic	97.80%	1999	[19]
GRNN	98.80%	2004	[20]
Fuzzy + KNN	99.14%	2006	[21]
Hybrid SVM	99.51%	2008	[22]

This paper reached, with the Bayesian Networks, an accuracy slightly higher than the ones presented in the first papers that used this dataset. However some advanced ma-

chine learning algorithms were developed and recent pappers reached levels of accuracy next to 100% classifing the instances in this dataset. Table X presents some of those papers and the accuracy that each one achived.

6. REFERENCES

1. IARC. World cancer report: International agency for research on cancer. Lyon, 2008.

2. NCI. SEER: Cancer Statistics Review. 2012.

3. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, Ransohoff DF: International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst 95(18):13841393, 2003.

4. Veronesi U, Boyle P, Goldhirsch A, Orecchia R, Viale G: Breast cancer. Lancet 365:17271741, 2005.

5. Joann G. Elmore, MD, MPH; Katrina Armstrong, MD; Constance D. Lehman, MD, PhD; Suzanne W. Fletcher, MD, MSc: Screening for Breast Cancer. The Journal of the American Medical Association, 2005.

6. Raimond W. M. Giard MD, Jo Hermans: The value of aspiration cytologic examination of the breast a statistical review of the medical literature. American Cancer Society, 2006.

7. Shweta Saxena, Kavita Burse: A Survey on Neural Network Techniques for Classification of Breast Cancer Data. International Journal of Engineering and Advanced Tecnology, 2012.

8. Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository.

9. Dataset Description. Available at: UCI Machine Learning Repository.

10. William H. Wolberg, W. Nick Street, O. L. Mangasarian: Machine Learning Techniques To Diagnose Breast Cancer rom Image- Processed Nuclear Features of Fine Needle Aspirates. National Science Foundation, 1993.

11. William H. Wolberg, W. Nick Street, O. L. Mangasarian: Breast Cancer Diagnosis and Prognosis Via Linear Programming. Journal of Operations Research, 1995.

12. William H. Wolberg, O. L. Mangasarian: Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. PNAS - Proceeding of the National Academy of Sciences, 1990.

13. William H . Wolberg, : Sparsity Through Automated Rejection. University College London, 2001.

14. Foster Provost: Machine Learning from Imbalanced Sets. New York University, 2000.

15. Manolis Maragoudakis, Katia Kermanidis, Aristogiannis Garbis, Nikos Fakotakis: Dealing with Imbalanced Data using Bayesian Techniques. University of Patras.

16. Ian H. Witten, Eibe, Frank, Mark A. Hall: Data Mining, Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2011.

17. Nitesh V. Chawla: C4.5 and Imbalanced Data sets: Investigating the efect of sampling method, probabilistic estimate, and decision tree structure. Workshop on Learning from Imbalanced Datasets II, 2003.

18. Kristin P. Bennett, O. L. Mangasarian: Neural Network Training via Linear Programming. Elsevier Science Publishers B. V., 1992.

19. Carlos Andres Peña-Reyes, Moshe Sipper: A Fuzzy-Genetic Approach to Breast Cancer Diagnosis. Elsevier, 1998.

20. Tüba Kiyan, Tülay Yildirim: Breast Cancer Diagnosis Using Statistical Neural Networks. Journal of Electrical & Eletronics Engineering, 2004.

21. Seral Sahan, Kemal Polat, Halife Kodaz, Salih Günes: A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine, V. 37, Pages 415-423, 2007.

22. Mehmet Fatih Akay: Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, V. 36, Pages 3240-3247, 2008.