Machine Learning Approaches to Automatic BI-RADS Classification of Mammography Reports

Bethany Percha

I. Introduction

The average American radiologist interprets at least 1,777 mammogram reports each year, or approximately one new mammogram every 70 minutes [1]. Because radiologists interpret so many mammograms and because the proper interpretation of a screening mammogram is often a matter of life or death for the woman involved, various attempts have been made to streamline the mammography reporting process and introduce consistent structure and terminology into mammography reports.

One important advance is the BI-RADS assessment coding scheme (Figure 1), a seven-level classification used to summarize a report and classify it into a distinct category based on the radiologist's overall assessment of the case. The BI-RADS assessment codes are designed to be translatable across physicians and institutions, and to serve as a basis for clinical follow-up. They also provide a convenient tool for researchers, since the codes are machine-interpretable and can be used in lieu of unstructured text-based diagnoses in large-scale clinical studies.

0 Incomplete
1 Negative
2 Benign finding(s)
3 Probably benign
4 Suspicious abnormality
5 Highly suggestive of malignancy
6 Known biopsy - proven malignancy

Fig. 1. The seven possible BI-RADS assessment codes and their meanings.

In theory, two radiologists who independently assess the same mammogram will produce reports with very similar terminology and identical BI-RADS codes. And in fact, if radiologists use consistent terminology to describe what they see, it should be possible to assign the BI-RADS codes automatically; a computer should be able to "read" the report and predict the code based on the radiologist's description of the image. A learning algorithm that could perform this task would be useful in three ways:

1. *Quality control.* The BI-RADS class represents the radiologist's overall impression of an image, and the rest of the report contains a description of what the radiologist saw. If two radiologists use very similar terminology to describe an image but assign it to different classes, it means that either the BI-RADS class boundaries are unclear or at least one of the radiologists was using insufficiently precise terms to describe the image.
2. *Training software.* A large part of radiology residents' training consists of learning to describe image features in a consistent way so that other physicians can easily interpret the report. A learning algorithm could be used to develop software that could provide real-time feedback to radiology residents (i.e. "Doctor, your description of this mammogram seems to correspond to BI-RADS class 3. Do you agree? If not, please modify your description.").
3. *Feature selection.* The algorithm could be used to help find the words and phrases most indicative of each class and make radiologists aware of the terms their colleagues are using to describe the various classes, thereby increasing consistency in reporting.

In this report, I describe a machine-learning approach to the automatic BI-RADS classification of mammography reports based on fairly extensive preprocessing followed by the application (and optimization) of several supervised learning techniques, including multinomial Naive Bayes, K-nearest-neighbors, and Support Vector Machines (SVM).

II. Preprocessing

A. Constructing the Training Corpus

A total of 41,142 mammography reports were extracted from the radTF database at Stanford University. The radTF database was designed by radiologists Bao Do and Daniel Rubin and serves as an educational tool for the training of radiology residents. Of these reports, 38,665 were diagnostic mammograms (not readings of mammograms from outside facilities, descriptions of ultrasound-guided biopsy procedures, or analyses of extracted tissue specimens). Of the diagnostic mammogram reports, 22,109 contained BI-RADS codes (older reports frequently do not have them) and were unilateral (single-breast) mammography reports. These 22,109 unilateral, diagnostic reports constituted the training corpus.

B. Report Structure

Mammography reports present an ideal corpus for work at the interface of machine learning and natural-language processing. Because of the large number of mammograms conducted each year, both in the United States and around the world, physicians have established fairly strict guidelines for the structuring of these reports. The reports usually consist of three sections:

history Begins with report type, such as "unilateral diagnostic mammogram". Patient's personal and clinical history, including age, cancer status, fam-

Department of Biomedical Informatics, Stanford University, Stanford, CA 94305–9510. E-mail: blpercha@stanford.edu. This report was submitted as my final project for CS 229 (12/9/10), Autumn 2010.

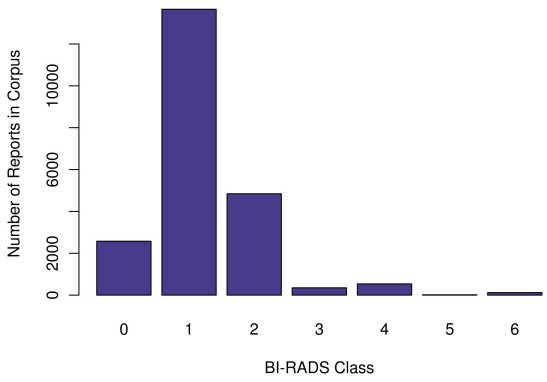

Fig. 2. Distribution of reports in training corpus by BI-RADS outcome class.

ily history, medications, and previous radiological exams/surgeries are described.

findings Includes a description of the breast tissue density and any defining features, such as scars from previous biopsies. Location and characteristics of any masses are described, including size, position, and nature of margins (edges).

impression Contains one or more BI-RADS assessment codes and the radiologist's impression of whether the patient's condition warrants additional clinical follow-up.

Most of the relevant description of the breast tissue and any abnormalities is in the "findings" section of the report, while the final BI-RADS classification is typically located at the beginning of the "impressions" section. This is extremely convenient from the standpoint of classifier training, since the "impressions" sections can easily be removed from the report and parsed to determine the BI-RADS assessment class (outcome variable), while the "findings" sections can serve as the input features for the classifier.

C. Assigning Outcome Classes

The BI-RADS outcome class was obtained for each report by searching the "impression" section for the term "bi-rads" or "birads" (not case-sensitive) and using the first number found after the term as the category. This method worked surprisingly well - out of 500 randomly-chosen reports confirmed by hand, 100% were classified correctly. The numbers of reports in each BI-RADS outcome category are shown in Figure 2. The vast majority of mammograms were negative for cancer.

D. Further Preprocessing

A significant amount of further preprocessing was required before any machine learning techniques could be tried. A summary of the preprocessing procedure is shown in Figure 3. One major problem with these reports was the amount of misspelled and concatenated words. Raw radiology reports are usually spoken aloud, recorded, and later transcribed into text by assistants. This can lead to many unusual typographical errors; a list of all the unique words

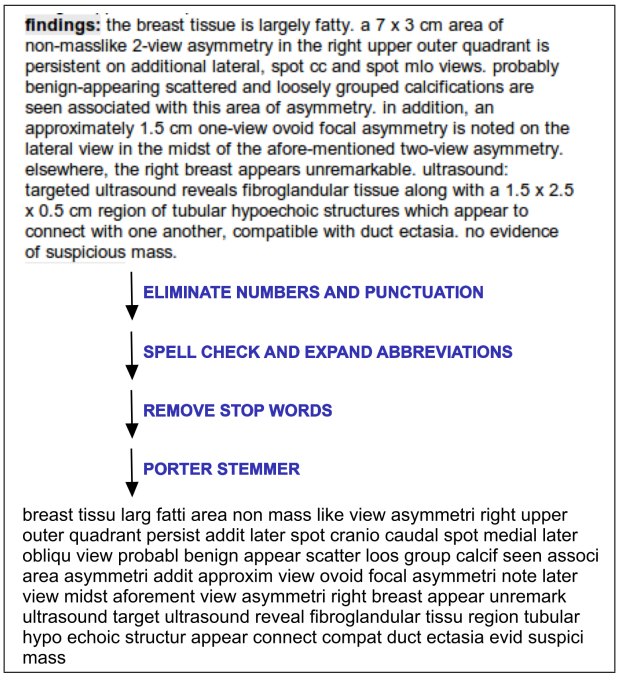

Fig. 3. Preprocessing pipeline for mammography reports. The processed "findings" sections were then converted into feature vectors with components equal to the number of times each stem appeared in the report.

found in the corpus (one per line) was a document over 160 pages long. Furthermore, many of these words were actually abbreviations or shorthand forms of longer words. For example, the terms "abnl", "abnormal", and a dozen or more misspellings and concatenations all referred to the same thing. I constructed my own spell-checker to deal with the misspellings and concatenations, and to expand the abbreviations, but the process was laborious.

Briefly, the preprocessing procedure consisted of first extracting the "findings" sections from the reports, eliminating numbers and punctuation, spell-checking the text, and expanding the abbreviations. Stop words (high-frequency words with little information content, such as "and" and "of") were removed, and a Porter stemmer [2] was applied so that variants such as "asymmetric" and "asymmetry" would appear as a single feature in the feature vector. The original and preprocessed versions of a sample report are shown in Figure 3. A corpus-wide histogram of the remaining word stems (ordered by frequency) is shown in Figure 4. All of the preprocessing for this project was done using custom scripts in Python (Version 2.6) [3].

E. Construction of Feature Vectors

After preprocessing, the entire corpus of "findings" sections consisted of patterns of 2,216 unique stems. The processed reports were then converted into feature vectors, where each element was the number of times a given stem appeared in a report. Interestingly, many of the reports used identical terminology; Figure 5 shows four "template" findings sections that together accounted for the majority (77%) of all BI-RADS class 1 reports.

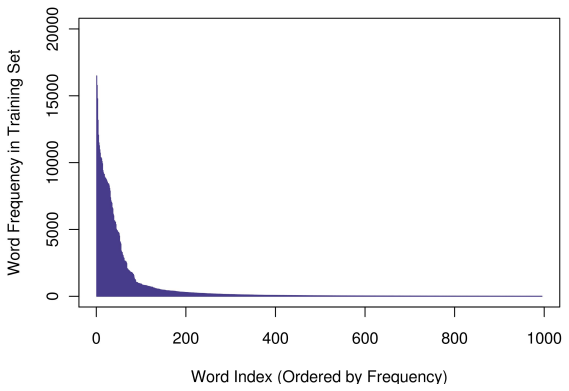

Fig. 4. Word stems ranked by frequency in the training corpus.

Freq.	Text
4002	"Breast tissue is of scattered fibroglandular density. There are no mammographic features of malignancy."
2832	"Breast tissue is heterogeneously dense which decreases mammographic accuracy. There are no mammographic features of malignancy."
2531	"Breast tissue is largely fatty replaced. There are no mammographic features of malignancy."
1178	"Breast tissue is dense which decreases mammographic accuracy. There are no mammographic features of malignancy."

Fig. 5. Four common "findings" sections for BI-RADS 1 reports. These exact combinations of words were used in over 77% of BI-RADS 1 reports.

III. Classification

A. Multinomial Naive Bayes and K-Nearest-Neighbors

Once the feature vectors were constructed, multinomial Naive Bayes and KNN (best performance: $K = 10$) classifiers were used to classify the reports using 10-fold cross-validation. The Naive Bayes and KNN parts of the analysis were done in Weka (Version 3.6.0) [4]. Standard Naive Bayes (which used binary instead ordinal features) achieved 76.4% cross-validation accuracy, multinomial Naive Bayes achieved 83.1% accuracy, and KNN achieved 83.0% accuracy. A learning curve for the multinomial Naive Bayes classifier showed that the misclassification error was the result of high bias (underfitting); the classifier performed just as well using only 10% of the training data, and its performance leveled out when larger training sets were used. As a result, the goal of subsequent analyses was to obtain a larger and more optimal set of features.

B. Support Vector Machines

With the goal of eventually moving to a higher-dimensional feature space, Support Vector Machines were also used to classify the reports. The first trials were performed using LIBLINEAR [5], which did not increase the dimensionality of the feature space (LIBLINEAR uses only a linear kernel) but did show that even in the original feature space, SVMs outperformed both Naive Bayes and KNN, achieving 89.0% classification accuracy on 10-fold cross-validation. LIBLINEAR offers two options for dealing with multiple outcome classes, one based on the standard L2-norm-penalized SVM and the other based on a sequential dual formulation by Crammer and Singer [6], but these showed no difference in performance. A grid search [7] was used to optimize the cost parameters for the SVMs; all results shown in Figure 6 are for the optimized classifiers.

C. Feature Weighting

In document classification, it is common to weight the features (words or word stems) using some measure of their importance. Usually this measure is based on term frequency in the corpus, but there are many options. Two forms of term frequency weighting were used in this project: standard TFIDF weighting and TFCNFX weighting, which is known to outperform TFIDF in some cases [8]. The weight of term i in document j is given by:

$$\textbf{TFIDF} \rightarrow \quad w_{ij} = \frac{\mathrm{tf}_{ij} \cdot \log(N/n_i)}{len(j)}$$

$$\textbf{TFCNFX} \rightarrow \quad w_{ij} = \frac{\mathrm{tf}_{ij} \cdot \log(N/n_i)}{\sqrt{\sum_{k=1}^{len(j)} \left(\mathrm{tf}_{kj} \cdot \log(N/n_k)\right)^2}}$$

where $len(j)$ is the length of document j, tf_{ij} is the number of times term i appears in document j, N is the total number of documents, and n_i is the number of documents in which word i appears. Standard TFIDF weighting actually decreased the performance of the SVM classifiers, although it increased the performance of the KNN classifier (Figure 6). TFCNFX weighting, on the other hand, always increased classifier performance. The linear SVM still performed best at this point, with 89.3% accuracy on 10-fold cross-validation.

D. Polynomial Kernel

Although the linear SVM performed well, transforming the data into a higher-dimensional feature space still held promise for reducing misclassification error. The degree-2 polynomial kernel is often used in text classification because its feature space is the space of all one- and two-word phrases contained in the document. The LIBLINEAR-POLY2 library is nearly as fast as LIBLINEAR and uses the degree-2 polynomial kernel; it was therefore employed to test whether transformation into a higher-dimensional feature space could improve performance. In fact, after optimization the polynomial SVM classifier performed nearly 1% better, achieving 90.1% accuracy on 10-fold cross-validation (see Figure 6).

E. Feature Selection

To obtain a set of features that would optimize the performance of the degree-2 polynomial classifier, the features were ranked using chi-squared attribute evaluation in Weka. Subsets of features of different sizes were then used

Technique	Percent Accuracy		
	No Weighting	TFIDF	TFCNFX
Naive Bayes	76.4		
Multinomial Naive Bayes	83.1		
K-Nearest Neighbors (K=10)	83.0	85.3	87.5
Support Vector Machines			
LIBLINEAR (L2-norm, one-against-one)	89.0	88.8	89.3
LIBLINEAR (Multiclass Crammer)	89.0	88.7	89.3
LIBLINEAR-POLY2 (polynomial kernel, degree 2)	88.8	88.6	90.1

Fig. 6. Comparison of 10-fold cross-validation accuracy for various classification techniques. The best-performing technique was a SVM (L2-norm, one-against-one for multiple outcome classes) that used a second-degree polynomial kernel.

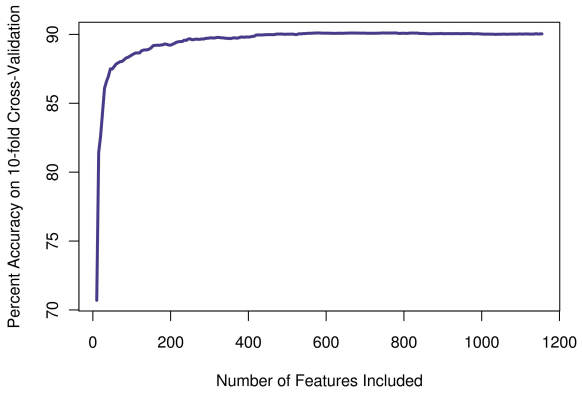

Fig. 7. Performance of degree-2 polynomial classifier vs. number of features used in classification. Features were initially ordered using chi-squared attribute evaluation.

to train the classifier and its performance was evaluated using 10-fold cross-validation. The results of this analysis are shown in Figure 7. The classifier achieves optimal performance using only the top 600 features, although performance does not decrease when additional features are included. Confusion matrices for the optimized classifier are shown in Figure 8.

IV. Most Informative Features

The list of the most informative features chosen by chi-squared attribute evaluation is interesting in its own right. The top 21 features are shown in Figure 9. The stem "incomplet" was most informative because it is seldom if ever used outside the context of BI-RADS class 0 (incomplete) reports. Although common, the stem "breast" was informative because it served as a proxy for the overall length of the report. Shorter reports tended to describe mammograms that were negative for cancer, while more serious cases warranted more description. Stems used to describe previous biopsies (such as "scar" and "stabl") helped distinguish reports from class 2, which often includes patients with fibrocystic breast disorders that lead to repeated biopsies for benign cysts. Terms referring to widespread or ill-defined masses or calcifications ("pleomorph", "calcif") or those used to localize masses within the breast ("oclock", "nippl") usually correspond to BI-RADS class 4, 5, or 6.

V. Some Informative Errors

Figure 8 shows that the classifier performed very well on certain classes of report and less well on others. It was excellent at splitting class 0 and 1 reports off from the others; this is likely because class 0 reports contained the telltale stem "incomplet" and class 1 reports often used identical terminology (Figure 5). The classifier also performed well on class 2 reports, though it sometimes mistook a report of a benign lesion as a fully-negative (class 1) report.

Class 3 reports ("probably benign") were only classified correctly 9.7% of the time, reflecting the inherent ambiguity of that class and the fact that it is, in some respects, the most dangerous class for a patient to be in. Many of the reports in class 3 used terminology that sounded as harmless as a class 2 or even class 1 report, which is why 49.1% of class 3 reports ended up in class 2 and another 21.1% ended up in class 1. However, 12.6% of class 3 reports used terminology that sounded more sinister than a description of a benign cyst. Some of this could be due to negation; for example, one misclassified report included the phrase, "no new focal dominant mass, architectural distortion, or suspicious microcalcifications are identified," which was full of class 4-6 terms once the stop word "no" was removed.

Class 5 reports were all classified as class 4, which again reflects the fuzzy conceptual boundary between "suspicious abnormality" and "highly suggestive of malignancy". In a way, class 5 is really a subset of class 4, and is assigned very infrequently, perhaps only to emphasize the fact that the patient needs clinical follow-up as soon as possible.

Perhaps the most interesting result of all was how frequently class 6 reports were misclassified (63.1% of the time). The reason appears to be that the patients involved have already been diagnosed with breast cancer and are usually undergoing treatment. The mammograms are used to monitor the course of treatment, and often the radiologist does not describe the cancerous lesions in full detail. In addition, patients who are in remission from breast cancer are still assigned to class 6, so some class 6 reports can actually sound quite optimistic.

VI. Conclusions and Future Directions

In conclusion, it is indeed possible to classify mammography reports into BI-RADS classes using a bag-of-words

Class	*Classified As...*						
	0	1	2	3	4	5	6
0	**2414**	60	79	2	21	0	1
1	48	**12789**	802	16	4	0	0
2	45	538	**4220**	7	29	0	5
3	25	74	172	**34**	44	0	1
4	46	20	57	3	**409**	0	4
5	0	0	0	0	18	**0**	0
6	6	6	30	1	34	0	**45**

Class	*Classified As...*						
	0	1	2	3	4	5	6
0	**93.7**	2.3	3.1	0.1	0.8	0.0	0.0
1	0.4	**93.6**	5.9	0.1	0.0	0.0	0.0
2	0.9	11.1	**87.1**	0.1	0.6	0.0	0.1
3	7.1	21.1	49.1	**9.7**	12.6	0.0	0.3
4	8.5	3.7	10.6	0.6	**75.9**	0.0	0.7
5	0.0	0.0	0.0	0.0	100.0	**0.0**	0.0
6	4.9	4.9	24.6	0.8	27.9	0.0	**36.9**

Fig. 8. Final confusion matrices for the LIBLINEAR-POLY2 classifier using the top 600 features. The left matrix includes absolute numbers of reports, while the right matrix shows the distribution of classifications (in percentages) for each real BI-RADS class. The diagonal elements of the matrices (correct classifications) are bolded.

Rank	**Stem**	**Most Common Context**	**Occurrences per report by class**						
			0	1	2	3	4	5	6
1	incomplet	*incompletely evaluated*	**0.9**	0.0	0.0	0.0	0.0	0.0	0.0
2	nippl	*x cm from the nipple* (Describing a mass.)	1.2	0.1	0.2	0.8	2.3	**4.6**	2.8
3	evalu	*incompletely evaluated*	**1.0**	0.0	0.0	0.1	0.1	0.1	0.2
4	breast	(Many contexts.)	4.2	1.9	3.8	4.6	5.7	6.9	**7.7**
5	featur	*no mammographic features of malignancy*	0.1	1.1	**1.2**	0.1	0.1	0.1	0.1
6	malign	*no mammographic features of malignancy*	0.1	1.1	**1.2**	0.1	0.1	0.3	0.3
7	neg	*breast is negative*	**0.8**	0.1	0.1	0.2	0.3	0.3	0.4
8	stabl	*stable post-biopsy change*	0.2	0.3	**1.5**	0.7	0.4	0.2	0.5
9	mammograph	*no mammographic features of malignancy*	0.3	1.5	**1.8**	0.7	0.9	1.7	1.2
10	calcif	*calcifications*	0.6	0.1	0.7	1.3	1.5	1.9	**2.0**
11	left	(Many contexts.)	1.2	0.4	1.2	1.8	2.3	3.4	**3.6**
12	marker	*scar marker* (post-biopsy marker)	0.2	0.3	1.3	0.8	0.9	0.7	**2.5**
13	right	(Many contexts.)	1.2	0.3	1.1	1.7	2.3	2.4	**3.1**
14	echoic	*hypoechoic mass*	0.0	0.0	0.1	0.5	1.2	**1.6**	1.3
15	biopsi	*post-biopsy change*	0.2	0.3	1.2	0.7	0.8	0.2	**2.9**
16	hypo	*hypoechoic mass*	0.0	0.0	0.1	0.3	0.9	**1.4**	1.2
17	oclock	(Describing mass location.)	0.2	0.1	0.2	0.9	2.1	**3.6**	2.7
18	post	*post-biopsy change*	0.2	0.3	1.2	0.7	0.7	0.7	**2.1**
19	scar	*scar marker* (post-biopsy marker)	0.2	0.2	**1.0**	0.4	0.3	0.2	0.2
20	pleomorph	*pleomorphic calcifications*	0.0	0.0	0.0	0.0	0.5	**1.2**	0.7
21	mass	(Many contexts.)	0.7	0.3	0.7	1.7	2.9	**5.2**	5.0

Fig. 9. The most informative features, along with the mean number of times they appeared in a report from a given class.

approach to the report text combined with a machine-learning algorithm based on an SVM in a high-dimensional feature space. Such an algorithm would probably prove most useful in the context of training software, especially for teaching residents how to unambiguously differentiate benign and suspicious lesions. It could also be useful in decision-support software, allowing radiologists to determine whether their descriptions and recommendations for clinical follow-up are consistent with those of hundreds of other physicians.

The most severe bottleneck in this project occurred at the preprocessing stage, and was the result of the large number of misspellings, concatenations, and abbreviations found in these reports. The homemade spell-checker used for this project is not a sustainable solution to this problem, since the number of errors appears to increase linearly with the size of the training set. String kernels could provide a potential solution to this problem [9], and may constitute the next phase of this project.

VII. Acknowledgements

I am extremely grateful to Bao Do and Daniel Rubin for helping me obtain the original mammography reports from the radTF database. Thanks guys!

References

[1] Smith-Bindman R, Miglioretti DL, Rosenberg R, et al, *Physician workload in mammography*, Am J Roentgenol, 190(2): 526-32, 2008.

[2] Han B, Obradovic Z, Hu ZZ, et al, *Substring selection for biomedical document classification*, Bioinformatics 22(17): 2136-42, 2006.

[3] Beazley D, *Python Essential Reference.* SAMS, 2009.

[4] Hall M, Frank E, Holmes G, et al *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, Volume 11, Issue 1, 2009.

[5] Fan RE, Chang KW, Hsieh CJ, et al *LIBLINEAR: A Library for Large Linear Classification*, Journal of Machine Learning Research, 9: 1871-4, 2008.

[6] Keerthi SS, Sundararajan S, Chang KW, et al *A sequential dual method for large scale multi-class linear SVMs.* Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.

[7] Hsu CW, Chang CC, Lin CJ (2010) *A practical guide to support vector classification.* From the LIBLINEAR website: http://www.csie.ntu.edu.tw/c̆jlin. Accessed 11/30/10.

[8] Salton G, Buckley C *Term-weighting approaches in automatic text retrieval.* Information Processing and Management Vol. 24, No. 5 pp. 513-523, 1988.

[9] Lodhi H, Saunders S, Shawe-Taylor J, et al *Text classification using string kernels.* Journal of Machine Learning Research, 2: 419-44, 2002.