Using BI-RADS Descriptors and Ensemble Learning for Classifying Masses in Mammograms

Yu Zhang, Noriko Tomuro, Jacob Furst, and Daniela Stan Raicu

College of Computing and Digital Media
DePaul University, Chicago, IL 60604, USA
{jzhang2,tomuro,jfurst,draicu}@cs.depaul.edu

Abstract. This paper presents an ensemble learning approach for classifying masses in mammograms as malignant or benign by using Breast Image Report and Data System (BI-RADS) descriptors. We first identify the most important BI-RADS descriptors based on the information gain measure. Then we quantize the fine-grained categories of those descriptors into coarse-grained categories. Finally we apply an ensemble of multiple Machine Learning classification algorithms to produce the final classification. Experimental results showed that using the coarse-grained categories achieved equivalent accuracies compared with using the full fine-grained categories, and moreover the ensemble learning method slightly improved the overall classification. Our results indicate that automatic clinical decision systems can be simplified by focusing on coarse-grained BI-RADS categories without losing any accuracy for classifying masses in mammograms.

Keyword: Mass Classification, BI-RADS, CADx.

1 Introduction

Breast cancer is the second leading cause of cancer related deaths for women in the U.S. after lung cancer [1]. At present, mammography screening is the most effective method for the early detection of breast cancer. However, the error rate of mammography screening is still high [2]. Many Computer-Aided Diagnosis (CADx) systems have been developed as a second opinion to assist radiologists [3].

Breast Image Report and Data System (BI-RADS) [4] is a set of lexicons describing breast lesions, which was developed by the American College of Radiology (ACR) to standardize the terminology in mammogram reports. The BI-RADS has been used in various research studies for the diagnoses of breast cancer. Kim et al. [5] used BI-RADS-based features, and applied a Support Vector Machine based on recursive feature elimination (SVM-RFE) for classifying abnormalities in mammogram images. Elter et al. [6] presented two CAD systems which use decision-tree learning and case-based reasoning for the prediction of breast cancer from BI-RADS attributes.

The research presented in this paper is part of an ongoing project for developing an image-based CADx system to classify suspicious masses in mammograms as malignant or benign. By studying BI-RADS descriptors, we will be able to identify

B. Caputo et al. (Eds.): MCBR-CDS 2009, LNCS 5853, pp. 69–76, 2010.

important domain knowledge and effective methods to guide our image-based CADx system. For radiologists, the shape and margin of masses are two important descriptors to distinguish malignant from benign masses [7]. Mass shape and margin feature are both defined by five categories in BI-RADS. In this research, we hope to simplify the decision process of classifying suspicious masses by using the coarse-grained BI-RADS categories, and still achieve equivalent or higher classification accuracies with the ensemble learning method. By applying the same methods from this study, the technical aspect involved in an image-based system could be simplified without sacrificing any classification accuracy.

Figure 1 below depicts the schematic framework of our approach. First the BI-RADS descriptors are extracted from the overlay files which contain keywords that describe each abnormality; next feature selection is applied to identify the important descriptors, and the fine-grained categories of those descriptors are collapsed and converted into coarse-grained categories; then the dataset is split into subsets based on the coarse-grained categories; finally, an ensemble of classifiers (Decision Tree, Bayes Network, Neural Network, Support Vector Machine, and K-Nearest Neighbor) is formed for each data subset, and the one which produced the highest accuracy is selected. The final classifications for the whole dataset are obtained by combining the classifications derived by the individual classifiers. The results showed that, by using the BI-RADS shape descriptor with coarse-grained categories along with the margin descriptor and patient age feature, our ensemble learning method achieved the overall accuracy of 84.43% for classifying masses in mammograms as malignant or benign.

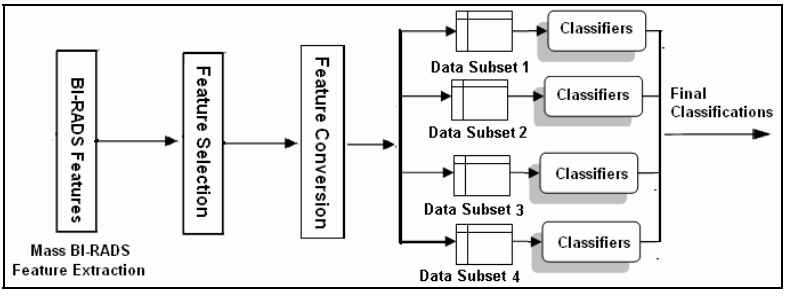

Fig. 1. Overall Framework of Our Approach

This paper is organized as follows: Section 2 reviews the BI-RADS descriptors and dataset, Section 3 describes the feature selection and data splitting methods, Section 4 presents the experimental results, and Section 5 discusses the conclusions and future work.

2 Data and BI-RADS Descriptors

2.1 Dataset Description

In this work, all mass instances were collected from the Digital Database for Screening Mammography (DDSM) from the University of South Florida [8], which is the

largest publicly available resource for the mammogram analysis research community. In DDSM images, suspicious regions of interest (ROIs; including masses and micro-calcifications) are marked by experienced radiologists, and BI-RADS information is also annotated for each abnormal region. In our experiment, we used mass instance images digitized by LUMYSIS. We removed instances with mixed BI-RADS descriptors or images with extreme digitization artifacts – that left us with a total of 681 mass instances, where 314 were benign and 367 were malignant. The BI-RADS descriptors were extracted from the overlay files using Matlab, and all classifications were conducted using a Machine Learning tool called Weka [9], with 10-fold cross-validation.

2.2 BI-RADS Descriptors

In BI-RADS, mass shapes are defined as either round, oval, lobulated, irregular or architectural distortion. Usually a poorly defined shape is more likely to be malignant than a well-circumscribed mass [7]. Margin is the border of a mass. In BI-RADS, five types of margins are defined: circumscribed, obscured, microlobulated, ill-defined and spiculated. Usually ill-defined margins or spiculated lesions are much more likely to be malignant [7]. Density is a description of the overall breast composition. In DDMS, the density value is between 1 and 4 where 1 means the breast is almost entirely fat, while 4 means the breast tissue is extremely dense [8].

The assessment descriptor in BI-RADS indicates the level of suspicion [8], which is a subjective interpretation and could vary among different radiologists. Since this descriptor cannot be computed or extracted from a mammogram image, the assessment descriptor was not used in our experiment.

In the experiment, we used four BI-RADS descriptors of masses: shape, margin, density, and patient age. The categories of the shape descriptor were converted into numeric values as: round = 1, oval = 2, lobulated = 3, irregular = 4 and architectural distortion = 5. Mass margin descriptors were also converted into numeric values as: circumscribed = 1, obscured =2, microlobulated = 3, ill-defined = 4 and spiculated= 5.

3 Methodology

3.1 Feature Selection

Feature selection is an important pre-processing step in classification. To obtain a good classification performance, it is critical to choose an optimal set of features for the given dataset [10]. To derive an optimal feature subset, we computed the Information Gain (IG) [11] for each descriptor. IG is a measure in Information Theory which indicates the informativeness of an attribute. Used in our context, IG essentially indicates the effectiveness of an attribute in the classification: a larger IG means the attribute is more informative. Gain(S,A) of an attribute A in a collection S is a measure based on Entropy:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where Value(A) is set of all possible values for attribute A, S_v is the subset of S with the attribute A of value v. And entropy [11] is a measure of purity, which can be used to indicate how pure a collection is. Entropy of a collection S is computed as:

$$Entropy(S) = -\sum_{j=1} p_j \log_2(p_j)$$

where p_j is the probability of the j subset in S (i.e., instances which belong to class j). Note that the IG value is computed for each attribute separately, and the resulting values are not dependent on the order of attributes selected.

Figure 2 shows the Information Gain of each of the four BI-RADS descriptors we investigated. Of them, margin has the largest information gain, which indicates that mass margin is probably the most important descriptor/feature for classifying masses. On the other hand, density has much lower information gain, which indicates that the feature could be nearly irrelevant for classification.

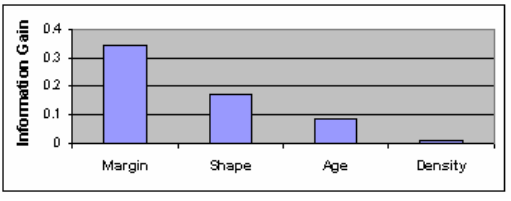

Fig. 2. Information Gain for each BI-RADS Descriptor

To verify this, we classified the dataset with and without the density descriptor and compared the performance with five different classifiers. Our experiment results (Table 1) show the classifications accuracies with density feature are slightly higher, however, the differences were not significantly for all five classifiers (p-value>0.05). Thus, we conclude that mass density descriptors were not important features, and can therefore be removed from the data without sacrificing classification performance.

Table 1. Classification Accuracies by Different BI-RADS Descriptors

BI-RADS / Classifiers	Margin, Shape, Age, Density	Margin, Shape, Age	P-Value
Decision Tree	83.99 %	83.26 %	>0.05
Bayes Network	79.74 %	79.74 %	>0.05
Neural Network	84.88 %	84.73 %	>0.05
SVM	82.38 %	82.09 %	>0.05
KNN	86.78 %	84.43 %	>0.05

3.2 Converting Features to Coarse-Grained Categories

The original BI-RADS shape descriptor has five categories (round, oval, lobulated, irregular and architectural distortion). Our motivation for grouping them into coarse-grained categories was to simplify the decision process of classification. To this end,

we first determined the splitting threshold by running the Decision Tree algorithm [11], where a decision tree was built using only the numeric shape feature with mass diagnosis as the target. Based on the decision tree, round, oval and lobulated shapes were categorized as "regular", and irregular and architectural distortion shapes were grouped as "irregular". In addition to the shape descriptor, we also converted the integer age feature into coarse-grained categories of "young" and "old". To determine the splitting threshold, we ran a Decision Tree and obtained the value of 57, in the same way as we did for the BI-RADS shape descriptor.

Note that, in our work in this paper, we converted the shape and age features to coarse-grained categories, but not the BI-RADS margin descriptor. That was because margin had by far the largest information gain over other features in our feature selection phase (as described earlier in section 3.1 above) – we assumed that finer categorization of this feature is critical in classifying masses.

3.3 Ensemble Learning and Partition Dataset into Subsets

Previous research has shown that ensemble learning often achieves better accuracy in classification than the individual classifiers that make them up [12]. In our work, the ensemble learning method partitions datasets and applies multiple classifiers as base classifications, and then combines the classifications from those partitioned datasets.

In our experiment, we tested three data partition schemes based on the features we converted to coarse-grained categories: 1) by age (young and old); 2) by shape (regular and irregular); 3) by age and shape (young regular, young irregular, old regular and old irregular). We expected that the overall classification accuracy could be improved by applying the best classification algorithms for each data subset. To find the best algorithms for a subset, we experimented with five classification algorithms: Decision Tree, Bayes Network, Neural Network, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). We chose those five algorithms because they have diverse characteristics. For example, Bayes Network is a statistical classifier; KNN decides the classification based on local information; Neural Network and SVM are known to be robust to noise. The final classification is the ensemble of multiple classifiers, where each classifier produces the highest accuracy among a data subset.

4 Experimental Results

First, we investigated the effect of converting the BI-RADS shape descriptor to coarse-grained categories. Table 2 below shows the accuracies by using the shape descriptor with coarse-grained categories, along with the margin descriptor and the age feature with original values. By comparing with the case using the original fine-grained shape categories, we can see that using the coarse-grained categories did not significantly decrease the accuracies for most classifiers (p-value > 0.05).

Table 2. Classification Accuracies for Fine vs. Coarse-grained Shape Categories

BI-RADS / Classifier	Shape (fine-grained) Margin, Age	Shape (coarse-grained) Margin, Age	P-value
Decision Tree	83.26 %	83.26 %	>0.05
Bayes Network	79.74 %	79.74 %	>0.05
Neural Network	84.73 %	83.26%	>0.05
SVM	82.09 %	81.64 %	>0.05
KNN	84.43 %	80.32 %	**<0.05**

Next we investigated the ensemble learning for the shape descriptor. Table 3 below shows the classification accuracies of the regular vs. the irregular shape subsets. Since the coarse-grained categories of the shape descriptor were used to partition the dataset, only the BI-RADS margin descriptor and the patient age feature were used for classification. Column (a) "Weighted Accuracy" in the table indicates the average accuracies weighted by the proportion of the size of the subsets. Column (b) is the classification accuracies using the BI-RADS margin descriptor and the age feature without any dataset partition. With shape partition scheme, the classification accuracies for the partitioned datasets had no significant difference compared with the classifications without data partition for most classifiers (p-value < 0.05).

Table 3. Classification Accuracies with Partitioned Datasets by Shape vs. without Partition

Classifier	Regular Shape 454 instances	Irregular Shape 227 instances	Weighted Accuracy (a)	No Dataset Partition (b)	P-Value
Decision Tree	81.50 %	86.78 %	83.26 %	82.53 %	>0.05
Bayes Net	79.74 %	86.78 %	82.09 %	81.20 %	>0.05
Neural Network	80.62 %	86.78 %	82.67 %	81.64 %	>0.05
SVM	81.06 %	86.78 %	82.97 %	81.20 %	>0.05
KNN	77.53 %	85.46 %	80.18 %	76.21 %	**<0.05**
Best Classifier	81.50 % Decision Tree	86.78 % Decision Tree	83.54 %	82.53 % Decision Tree	>0.05

Then we investigated the effect of converting the age feature to coarse-grained categories. Table 4 shows the classification accuracies of the young vs. old subsets. All classifications included three features: the coarse-grained categories of the BI-RADS shape descriptor, the BI-RADS margin descriptor and the patient age. Column (a) "Weighted Accuracy" in this table is calculated in the same way as the previous table. Column (b) is the classification without data partition. For this partition scheme, the classification accuracies had no significant difference compared with the classifications without data partition for all classifiers (p-value < 0.05).

Finally we investigated the ensemble learning on four data subsets partitioned by age and shape. All classifications used only two features: the BI-RADS margin descriptor and the age feature. Table 5 shows the classification accuracies. Column (a) "Weighted Accuracy" is calculated in the same way as the previous tables. Column

Table 4. Classification Accuracies with Dataset Partition by Age vs. without Dataset Partition

Classifier	Young Age 348 instances	Old Age 333 instances	Weighted Accuracy (a)	No Dataset Partition (b)	P-Value
Decision Tree	85.63 %	81.08 %	83.41 %	83.26 %	>0.05
Bayes Net	81.90 %	81.68 %	81.79 %	79.74 %	>0.05
Neural Network	84.77 %	81.38 %	83.11 %	83.26 %	>0.05
SVM	81.61 %	79.88 %	80.76 %	81.64 %	>0.05
K NN	83.05 %	78.98 %	81.06 %	80.32 %	>0.05
Best Classifier	85.63 % Decision Tree	81.68 % Bayes Net	84.00 %	83.26 % Decision Tree	>0.05

Table 5. Classification Accuracies of Datasets Partitioned by Age and Shape

	Younger Age Regular Shape	Younger Age Ir-regular Shape	Older Age Regular Shape	Older Age Irregular Shape	Weighted Accuracy With Partition (a)	Accuracy Without Partition (b)	P-Value
Classifier	273 instances	75 instances	181 instances	152 instances	681 instances	681 instances	
Decision Tree	87.55 %	85.33 %	76.80 %	87.50 %	84.43%	82.53 %	>0.05
Bayes Net	85.35 %	85.33 %	72.93 %	87.50 %	82.53%	81.20 %	>0.05
Neural Network	84.62 %	82.67 %	75.69 %	87.50 %	82.67%	81.64 %	>0.05
SVM	82.42 %	85.33 %	76.24 %	87.50 %	82.23%	82.33%	>0.05
KNN	80.22 %	82.67 %	76.80 %	81.58 %	79.88%	76.21 %	**<0.05**
Best Classifiers	87.55 % (Decision Tree)	85.33 % (Decision Tree and others)	76.80 % (Decision Tree)	87.50 % (Decision Tree and others)	84.43% (*)	82.53 % Decision Tree	>0.05

* 84.43% *is the weighted accuracy computed from the best classifiers of the last row.*

(b) is the classification accuracies without dataset partition. The ensemble learning with weighted classification accuracy achieved better performance over the best classification with no data partitioning (84.43% vs. 82.53%). Note that, the weighted accuracy is largely dragged by the low accuracy from the older age and regular shape group, where three other groups have achieved significantly better classifications. To achieve significantly better classifications with the ensemble learning, we will further investigate and improve the classifications of the older age regular shape group in our future work.

5 Conclusions and Future Work

In this paper, we explored an ensemble learning of using quantized BI-RADS features for classifying masses in mammograms. Our experiment showed that mass density descriptor could be removed without sacrificing classification performance. Using the

coarse-grained shape categories along with the margin descriptor and patient age, our ensemble classifier achieved the overall accuracy of 84.43%. Our results indicate that automatic clinical decision systems can be simplified by focusing on coarse-grained shape BI-RADS categories without losing any accuracy for classifying masses in mammograms.

In this experiment, we found that the mass instances of old age regular shape group produced much lower classification accuracies than other groups. This result suggests that the mass instances in this group are more difficult to classify, and using the age and the mass margin descriptor may not be enough to distinguish malignant from benign.

In future work, we are planning to apply the methods learned in this content-based classification system to build an image-based CADx system. And as our study indicates that margin is the most important feature for classifying masses, we plan to use even finer categorization such as continuous values to represent the margin feature in the image-based CADx system.

References

1. National Cancer Institute, American Cancer Society Cancer Facts & Figures (2008), http://www.cancer.org
2. Yankaskas, B.C., Schell, M.J., Bird, R.E., Desrochers, D.A.: Reassessment of Breast Cancers Missed During Routine Screening Mammography: A Community-Based. American Roentgen Ray Society 177, 535–541 (2001)
3. Cheng, H.D., Shi, X.J., Min, R., Hu, L.M., Cai, X.P., et al.: Approaches for Automated Detection and Classification of Masses in Mammograms. Pattern Recognition (2006)
4. D'Orsi, C.J., Bassett, L.W., Berg, W.A.: Breast Imaging Reporting and Data System: ACR BI-RADS-Mammography (ed 4). American College of Radiology, Reston, VA (2003)
5. Kim, S., Yoon, S.: BI-RADS Feature-Based Computer-Aided Diagnosis of Abnormalities in Mammographic. In: 6th International Special Topic Conference on ITAB (2007)
6. Elter, M., Schulz-Wendtland, R., Wittenberg, T.: Prediction of Breast Biopsy Outcomes Using CAD Approaches That Both Emphasize an Intelligible Decision Process. Medical Physics 34(11) (2007)
7. Winchester, D.J., Winchester, D.P., Hudis, C.A., Norton, L.: Breast Cancer, 2nd edn. Springer, Heidelberg (2007)
8. The Digital Database for Screening Mammography, http://marathon.csee.usf.edu/Mammography/Database.html
9. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
10. Kohavi, R., John, G.: Wrappers for Feature Subset Selection. Artificial Intelligence 97(1-2), 273–324 (1997)
11. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (2001)
12. Opitz, D., Maclin, R.: Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research 11, 169–198 (1999)