Review of Feature Selection Algorithms for Breast Cancer Ultrasound Image

Kesari Verma[1], Bikesh Kumar Singh[2], Priyanka Tripathi[3], and A.S. Thoke[4]

[1] Department of Computer Applications
National Institute of Technology, Raipur India
kverma.mca@nitrr.ac.in
[2] Department of Biomedical Engineering,
National Institute of Technology Raipur
bksingh.bme@nitrr.ac.in
[3] Department of Computer Application,
National Institute of Technology Raipur
ptripathi@nitttrbpl.ac.in
[4] Department of Electrical Engineering,
National Institute of Technology Raipur
asthoke.ele@nitrr.ac.in

Abstract. Correct classification of patterns from images is one of the challenging tasks and has become the focus of much research in areas of machine learning and computer vision in recent era. Images are described by many variables like shape, texture, color and spectral for practical model building. Hundreds or thousands of features are extracted from images, with each one containing only a small amount of information. The selection of optimal and relevant features is very important for correct classification and identification of benign and malignant tumors in breast cancer dataset. In this paper we analyzed different feature selection algorithms like best first search, chi-square test, gain ratio, information gain, recursive feature elimination and random forest for our dataset. We also proposed a ranking technique to all the selected features based on the score given by different feature selection algorithms.

Keywords: Feature Selection, Random Forest, Ranking of features, important feature selection.

1 Introduction

Correct classification of patterns in breast cancer is one of the important research issues of current era. Cancer is the second leading cause of death in developed countries and the third leading cause of death in developing countries. Ultrasound facility is one of the economic ways for early detection and screening.

Pathological tests are the most reliable and most traditional methods for disease detection. Computation in pathology data creates a revolution in the field of biological data. Fuchs and Buhmann[1] has given a concise definition for computational pathology.

D. Barbucha et al. (eds.), *New Trends in Intelligent Information and Database Systems,*
Studies in Computational Intelligence 598, DOI: 10.1007/978-3-319-16211-9_3

Computational pathology investigates a complete probabilistic treatment of scientific and clinical workflows in general pathology, i.e. it combines experimental design, statistical pattern recognition and survival analysis within an unified framework to answer scientific and clinical question in pathology [1]. In this paper we used ultrasound images dataset instead of using pathological data. Computer aided design helps to predict the correct class of cancer without an expert. The layout of breast cancer classification techniques is shown in Fig. 1. The figure consists of three parts. Part one shows the pathological investigation of tumor by medical experts in terms of size, randomness, ratio of height and width, these feature extraction requires human experts and it is time consuming process. Part two shows different computer aided techniques that used to extract the important features that contribute in image analysis and classification task. Our approach is to map the pathological features with computer aided features extracted from images. Part III is set of classification techniques that can be applied in order to accurately classify the images based on extracted feature from part II. The proportion of the total number of predictions that were correct (supervised learning) is called accuracy of classifier. Important recent problems in medical diagnosis is that images containing many input variables (hundreds or thousands), with each one containing only a small amount of information, identification of most important features is still challenging. A single feature selection algorithm will then have accuracy only slightly better than a random choice of classes, it is not able to select all the important feature that are contributing in classification process. Combining the random features can produce improved accuracy [10]. The feature selection techniques are applied from part II to Part III, in order to reduce computational time and space in memory.

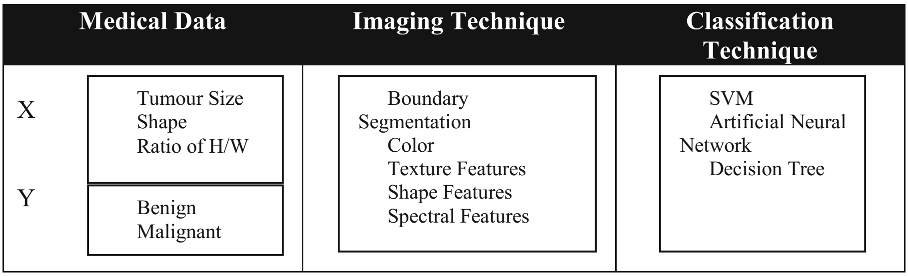

Fig. 1. Layout of Breast cancer Classification

1.1 Dataset Description

In this study we collected Ultrasound images data from J.N.N. Govt Hospital Raipur of Chhattisgarh, India. The images are labeled by medical professionals to train the model. Example of a benign and a malignant tumor is shown in Fig. 2. The geometric features of the images are the height, width, ratio of height and width, closeness of boundary.

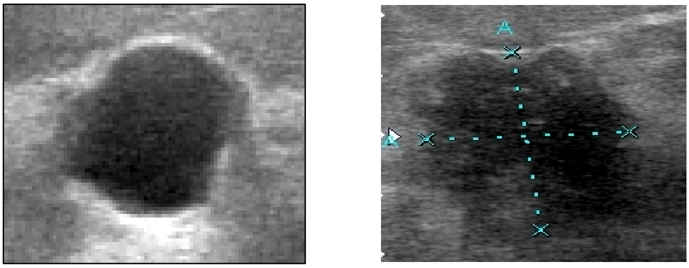

Fig. 2. (a) Benign (b) Malignant

2 Features

Features are defined as a function of one or more measurements, the values of some quantifiable property of an object, computed so that it quantifies some significant characteristics of the object. A set of features that helps the model to recognize the pattern is called class label. The feature set may contain set of irrelevant features. The irrelevant input features will induce great computational cost. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. The reduction of dimensionality the data and may allow learning algorithms to operate faster, accurately more effectively. The objective of this paper is find the most relevant features in order to discriminate different class of tumors in dataset. Table 1. shows different features extracted from ultrasound images.

Table 1. Extracted Features from Ultrasound Images

Particular	Number of Features
First Order Statistics (FOS) (features 1-5) [mean,variance,median,mode,skewness,kurtosis,energy,entropy]	06
Haralick Spatial Gray Level Dependence Matrices [2]	26
Gray Level Difference Statistics (GLDS) [contrast, entropy, energy, mean]	04
Neighbourhood Gray Tone Difference Matrix (NGTDM)	05
Statistical Feature Matrix (SFM) [feat, coarse, cont, period, roughness]	04
Laws Texture Energy Measures (TEM)	06
Fractal Dimension Texture Analysis (FDTA)	04
Shape (area, perimeter, perimeter^2/area)	04
Spectral Features	379
Law Features	06
Others	13
Total	457

2.1 Feature Selection Techniques

In previous section we described all 457 features extracted from ultrasound image. The feature selection process is divided into two categories feature ranking and subset selection which is shown in Fig. 3.

Feature Ranking. Kohavi and John [4] proposed variable ranking method for ranking the features based on their importance. Algorithm 1. demonstrate the layout of the feature ranking algorithm.

```
Algorithm 1. Ranking features
Input  : S ← set of features
Output : N ← Top n ranked features
Method
     1. Features ← Evaluation_criteria(D)  // Evaluation
        criteria  on  that  basis  the  features  are
        evaluated.
     2. Rank_features ← sort_descending(Features)
Return(top n ranked_features)
```

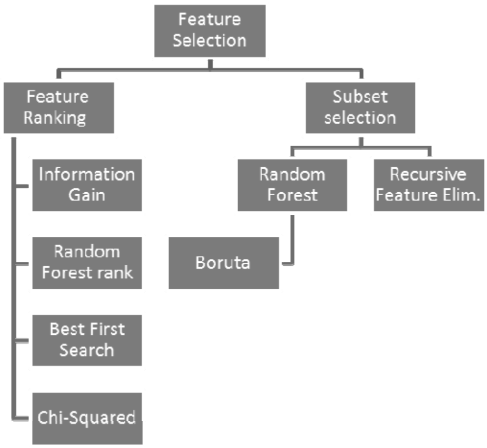

Fig. 3. Layout of different Feature Selection Techniques

Information gains, Gain Ratio, Best First search algorithm, Chi-Square test are some specific techniques that are widely using for feature selection purpose. The details are given as below.

Information Gain. This technique is based on decision tree induction ID3 [5] it uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon from information theory. If p_i represents the number of times tuples occurred in data D. This attribute minimize the information needed to classify the tuples in the resulting partitions. The information gain is represented by equation 1.

$$\inf(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Splitting attribute measures, that define information needed to exact classify the data is defined by equation 2.

$$\inf_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \inf(D_j) \tag{2}$$

Information gain is difference between original information and information after splitting is defined in equation 3.

$$Gain(A) = \inf(D) - \inf_A(D) \tag{3}$$

In this technique the features which have highest information will be ranked high otherwise low. Using Quinlan C4.5 algorithm [4] the attribute that are in the higher level of the tree are considered for further classification and these features have more importance.

Gain Ratio. C4.5 [5] a successor of ID3[6] uses, an extension to information gain known as gain ration. It applies a kind of normalization to information gain using split information defined in equation 5.

$$splitInfo(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{(|D_j|)}{(\lfloor D \rfloor)} \tag{4}$$

The gain ratio can be defined by equation 5. Intrinsic information: entropy of distribution of instances into branches by using equation 4.

$$GainRatio(S,A) = \frac{Gain(S,A)}{IntrinsicInfo(S,A)}. \tag{5}$$

Random Forest Filter. Breiman et. al [7] has proposed random forest algorithm, it is an ensemble approach that work as form of nearest neighbor predictor. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator [15]. Ensembles are divide-and-conquer tree based approach used to improve performance of classifier. The ensemble method is that a group of weak learners that group together and work as a strong learner to take the decision for unknown attributes. Giger's et. al. [16] developed a method for automated feature analysis and classification of malignant and benign clustered for micro calcifications data.

Best First Search. Best first search [14] is an Artificial Intelligence search technique which allows backtracking in search path. It is a hill climbing, best first search through the search space by making change in current subsets.

2.2 Feature Subset Selection

In this approach subsets of features are selected, subset feature selection is an exhaustive search process. If data contain N initial features there exist 2^N possible subsets. Selection of features from 2^N possible subsets is an exhaustive search process that is call heuristic search algorithm.

Subsets of features are selected and analyzed their classification accuracy, if it is increasing, that feature is selected otherwise rejected, a new set of feature are participate in evaluation process.

Many feature selection routines used a wrapper approach [4] to find appropriate variables such that an algorithm that searches the feature space repeatedly fits the model with different predictor sets. The best predictor set is determined by some measure of performance. The objective of each of these search routines could converge to an optimal set of predictors. The layout of subset feature selection method is shown in Algorithm 2.

```
Algorithm 2: Subset feature selection
S ← All subsets {}
For each subset   s ∈ S
      Evaluates (s)
Return {subset}
```

2.3 Recursive Feature Elimination [3]

Recursive feature elimination method is based on the concept that the features are eliminated recursively till the optimal set of features are not selected from the whole set. Random forest, forward subset selection, backword subset selection algorithm using caret, Boruta are the well-known techniques in R [8].

3 Experimental Study

For experimental evaluation of the proposed detection feature, we make use of our own created database of 188 ultrasound breast cancer images. The images were never-compressed gray scale images with resolution of 300x 340 and 90horizontal and vertical DPI. All feature extraction experiments were performed in MATLAB 2012 in windows environment with 4 GB RAM and 500GB Hard disk. For feature selection process we used R statistical package using caret [8], Borutha[9]. The selected features are with different feature selection algorithms is shown in Table 2.

Table 2. Selected Features using different techniques

Technique	Feature	
Information Gain	A50 100.00,A97 75.28 A137 69.66,A26 37.08 A93 22.47	Information from highest to lowest
Random Forest Rank	A33 6.325152, A53 -5.902789 A38 5.706619, A30- 5.491 A20 5.394731, A19-5.195 A50 5.162865, A42-5.158 A52 5.051218, A44-4.914 A34 4.883433, A171-4.714	Ranking of Features from top to bottom
Ranking features by Chi-Square	A141 A140 A171 A149 A163 A154 A137 A155 A158 A168 A169 A174 A175 A157 A172 A20 A55 A148 A152 A180 A165 A178 A156 A170 A177	
BestFirst Search	A33 A161 A170 A193	Exhaustive Search process stat with one feature and continue till the optimal features are not selected
Ranking by gain ratio	A163 100.000,A26 52.563 A33 39.641,A50 28.109 A44 23.424,A91 12.740 A8 11.594,A1 8.281	% of contribution from top to bottom
Random forest Feature selection measure [Borutha]	A20 A30 A33 A34 A37 A38 A42 A44 A50 A52 A53 A55 A171	
Recursive Feature Elimination	A50, A52, A39, A53, A17, A34, A35, A47, A49, A16, A32, A10	

In Table 2 it shows that information gain attribute selection measure selected 5 attribute A50, A137, A26 and A93. All the attributes are ranked based on information gain from higher to lower. Random forest rank selection algorithm ranked all feature from 6.32 to 4.71. Random forest rank algorithm and decision tree information gain selected 95% features from texture categories and 5% from spectral feature categories. Chi-square algorithm selected 23 important features from total 457feaures. Best first search is an exhaustive search algorithm that selected four features as important category. In Chi–Square and Best first search algorithm most of the features from spectral and very few features from texture categories. Gain ratio algorithm selected 8 features in important out of 7 are texture feature and one Spectral feature. Recursive feature elimination process selected all texture feature (12) are important feature but algorithm found spectral features are not important for classification. For Random forest feature selection 12/13 feature from texture and one feature selected from spectral category.

Based on Table 2 all features scores are evaluated which is shown in table 3. Table contain 0 and 1 value. 1 represents voted by algorithm 1 represent not voted by any of the algorithm. In experiment we found that Attribute A50 got highest ranked, six feature selection algorithm voted feature A40. We can conclude that Attribute 50 is most relevant feature for our dataset. Similarly other features are ranked attribute 20 has the lowest rank 2.

Table 3. Ranked Feature based on score of different feature selection algorithms

Features	Information Gain	Random Forest Rank	Chi Square	Best First Search	Gain Ratio	Random Forest	RFE	Rank
A50	1	1	0	1	1	1	1	6
A20	0	1	0	1	0	1	1	4
A33	0	1	1	0	1	1	0	4
A38	0	1	0	0	0	1	1	3
A42	0	1	0	0	0	1	1	3
A44	0	1	0	0	1	1	0	3
A52	0	1	0	0	0	1	1	3
A53	0	1	0	0	0	1	1	3
A55	0	0	0	1	0	1	1	3
A171	0	1	0	1	0	1	0	3
A19	0	1	0	0	0	0	1	2
	1	0	0	0	1	0	0	2
	0	1	0	0	0	1	0	2

All the selected features were evaluated using support vector classifier. The Classification accuracy using different kernel is shown in Table 4. The performance for kernel radial is highest in compared to other kernel so we selected radial kernel for measuring classification accuracy using selected features.

Table 4. Classification Accuracy using kernel gamma = .001 cost =10

Kernel	svm	tune svm
Radial Kernel	82.75	86.20
Polynomial	65.51	62.06
Gaussian	48.27	86.20
Sigmoid	62.06	79.31

3.1 Accuracy

Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications (quantity) divided by the total number of classifications [9]. Other parameters that also important for correct prediction and selection of classifier are precision, recall/sensitivity, specificity and F-measrue. Experimental result with all feature selected algorithm, its accuracy, accuracy after tuning svm parameters, True positive (TP), False Positive (FP), False Negative (FN), True Negative (TN), recall, precision, sensitivity and F- measures results are shown in Table 5.

Table 5. Accuracy Measurement parameters

Precision	$\frac{TP}{TP+FP}$
Recall/Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TP}{TN+FP}$
F-Measure	$2*\frac{\Pr ecision*\mathrm{Re}\,call}{\Pr ecision+\mathrm{Re}\,call}$

The classification accuracy for support vector machine is shown in first row, which is using all 457 feature. The accuracy and other parameter are also shown in respective rows. Principal component feature selection measure (dimension reduction) algorithm performs worst using svm classifier for our dataset. We applied our scored feature table that is named as hybrid feature in last column of the table it accuracy is as same as original svm classifier as well as f-measure value is also equal to original classifier value. Our hybrid feature selector selected 43 features as important for classification task. We can conclude that the instead of taking 457 feature only 43 features are highly contributing most relevant feature for breast cancer classification using ultrasound images. The experimental results are shown in Table 6.

Table 6. Experimental results

Feature selection Technique	Selected Attributes	Accuracy	After Tuning	Recall	Sensitivity	Precision	F-Measures
SVM	457	82.75	86.2	0.77	1	0.73	0.87
C5.0 Ranking	5	79.31	75.86	0.68	0.92	0.6	0.78
Random Forest	25	79.31	82.75	0.76	0.92	0.733	0.83
Chi Square Ranking	25	79.31	79.31	0.75	0.85	0.733	0.79
Breath First Search	4	75.86	72.41	0.65	0.92	0.533	0.76
Borutha	13	75.86	82.75	0.81	0.85	0.533	0.83
RFE	14	82.75	82.75	0.8	0.85	0.8	0.82
PCA	2 PC	68.96	51.72		0		NA
Hybrid		86.2	86.2	0.77	0.2	1	0.87

4 Conclusion

In this paper different feature selection algorithm were analyzed for ultrasound breast cancer images. We extracted 457 feature features from images of texture and spectral categories. We performed the experimentation with all 457 features using support vector machine using 10 fold cross validation. After tuning the parameter we achieved 86.2069% of classification accuracy. We applied feature selection algorithms and svm classifier using 10 cross fold validation to all 7 feature selection algorithms and principal component for dimension reduction. We created a score matrix based upon all selected features and voting of all algorithms. Based on the score we arranged the features in the descending order, with this technique 43 feature were selected and svm classification technique was applied for classification purpose. We achieved the same accuracy as we got for all 457 features. It also reduced the computational time and memory for classification purpose.

References

1. Thomas, J.F., Buhmann, M.J.: Computational pathology: Challenges and promises for tissue analysis. Computerized Medical Imaging and Graphics 35(7-8), 515–530 (2011)
2. Haralick, R.M., Shanmugam, K., Dinstein, M.I.: Texture Feature for Image Classification. IEEE Transaction on Systems, Man and Cybernetics 3(6), 610–619 (1973)
3. (July 2014), http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Dimensionality_Reduction/Feature_Selection
4. Kohavi, J.G.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997), doi:10.1016/S0004-3702(97)00043-X
5. Quinlan, J.R.: C4.5: Programs for Machine Learning. Machine Learning, vol. 16, pp. 235–240. Academic Kluwer Academic Publishers, Boston (1994)
6. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
7. Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. Journal of Statistical Software 36(11), 1–13 (2010)
8. R Statistical Package (July, 2014), http://CRAN.R-project.org/package=varSelRF
9. Svetnik, V., Liaw, A., Tong, C., Wang, T.: Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli, F., Kittler, J., Windeatt, T. (eds.) MCS 2004. LNCS, vol. 3077, pp. 334–343. Springer, Heidelberg (2004)
10. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
11. Davis, J., Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. Technical report #1551, University of Wisconsin Madison (January 2006)
12. Tan, P.N., Kumar, V., Steinbach, M.: Introduction to Data Mining. Pearson education, 321321367th edn. Addison-Wesley (2005) ISBN : 0321321367
13. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Ph.D. thesis in Computer Science. University of Waikato, Hamilton, New Zealand (1999)
14. Rich, E., Knight, K.: Artificial Intelligence. McGraw-Hill (1991)
15. Ensemble method (Ocober, 2014), http://scikit-learn.org/stable/modules/ensemble.html#b2001
16. Gruszauskas, N.P., Drukker, K., Giger, M.L., Chang, R.F., Sennett, C.A., Moon, W.K., Pesce, L., Breast, U.S.: computer-aided diagnosis system: robustness across urban populations in South Korea and the United States. Radiolog 253, 661–671 (2009)