Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Genetic Programming and Feature Selection for Classification of Breast Masses in Mammograms

R J Nandi[1], A K Nandi[1], R Rangayyan[2], and D Scutt[3]

***Abstract*— A dataset of 57 breast mass mammographic images, each with 22 features computed, was used in this investigation. The extracted features relate to edge-sharpness, shape, and texture. The novelty of this paper is the adaptation and application of genetic programming (GP). To refine the pool of features available to the GP classifier, we used five feature-selection methods, including three statistical measures – Student's t-test, Kolmogorov-Smirnov Test, and Kullback-Leibler Divergence. Both the training and test accuracies obtained were above 99.5% for training and typically above 98% for testing.**

I. Introduction

BREAST cancer is the most common cancer in England and Wales and the most common cause of cancer death in women [1]. Early detection is a key factor in prognosis and consequently plays a major role in reducing mortality and is currently performed by radiologists using mammography, with a significant human element needed for to the diagnosis. Techniques [2], [3], [4] are being developed to introduce computer-aided diagnosis (CAD) procedures for efficient screening and detection of breast cancer. Rangayyan et al. [5] proposed the use of shape factors and a measure of edge-sharpness known as acutance for the classification of manually segmented masses as benign or malignant, and as spiculated or circumscribed. They obtained an overall accuracy of 95% with a database of 54 mammographic images. A problem encountered in such studies lies in the selection of the best subset of features so as to facilitate efficient pattern classification. Sahiner et al. [6] studied this problem in the context of CAD of breast cancer. In this paper, a novel technique called genetic programming (GP) is introduced and adapted for classification of breast masses into the benign and malignant categories.

II. Data and Features

The data were extracted from mammograms obtained from Screen Test: Alberta Program for the Early Detection of Breast Cancer [7], with 37 regions of interest (ROIs) related to benign masses and 20 ROIs related to malignant tumors [8]. The images were digitized to a resolution of 50 μm with 12 bits per pixel; however, texture features were extracted after resampling to 200 μm and requantization to 8 bits per pixel. The diagnosis of each case was proven by biopsy. Benign and malignant ROIs were manually identified, and contours were drawn by a radiologist experienced in screening mammography.

Benign masses generally possess smooth and round contours, whereas malignant tumors typically exhibit rough contours with spiculations and concavities. Also, benign masses generally have homogeneous internal texture and sharp or well-circumscribed margins, whereas malignant tumors typically exhibit heterogeneous texture and ill-defined or blurred margins. A set of 22 features was extracted for each ROI [8]. The features include four edge-sharpness measures, four shape factors, and 14 statistical texture features.

III. Genetic Programming

Genetic programming (GP) [9] is a type of evolutionary learning algorithm and is an extension of the genetic algorithm (GA). The main difference between GP and GA lies in the representation of the solution. GP evolves computer programs as the solution, whereas GA creates a string of numbers or parameters that influence the performance of a fixed solution. Unlike GA, GP naturally allows nonlinear mapping and implicitly includes feature selection. GP is a relatively new classification technique, and has been successfully applied to a variety of classification problems [10], [11], [12]. GP embraces some concepts present in natural selection.

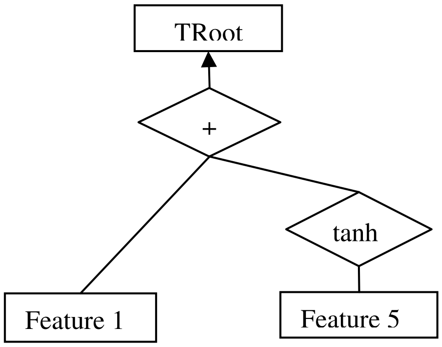

Fig. 1. An example of a GP tree.

[1] Department of Electrical Engineering and Electronics, The University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, UK(phone: +44-151-794-4525; fax: +44-151-794-4540; e-mail: a.nandi@liverpool.ac.uk).

[2] Department of Electrical & Computer Engineering, Schulich School of Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4 (e-mail: ranga@enel.ucalgary.ca).

[3] School of Health Sciences, The University of Liverpool, Thompson Yates Building, Liverpool, L69 3GB, UK (e-mail: dunc@liverpool.ac.uk).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

An outline of how GP works is given below. At the beginning, GP has available a set of feature values, a set of functions, and a set of parameters. Initially, a set of individuals, called GP trees, are created (randomly or otherwise) (for an example, see Figure 1). The idea is that one puts in the relevant feature values to a GP tree and gets out a number close to 0 or 1 depending on whether the condition is benign or malignant, respectively. The initial set of solutions (individuals) represents the first generation. The number of solutions in each generation is the population size.

GP uses the following steps:

- Create initial population
- Loop
    - Fitness evaluation of each individual
    - Selection of individuals
    - Modification (by GP operators):
        - Crossover
        - Mutation
        - Reproduction
- Until some criterion is met.

Three processes are employed to create the population for the next generation:

1) Crossover – Two new individuals are created by randomly selecting nodes from each parent and swapping these; see Figure 2.

2) Mutation – A node is randomly selected, the tree downstream from it is deleted, and a new sub-tree is generated from this node in exactly the same way as the initial population was grown; see Figure 3.

3) Reproduction – No change is made and an individual is simply copied.

These processes continue up to a certain number of generations. The final solution is the tree with the best score in the last generation.

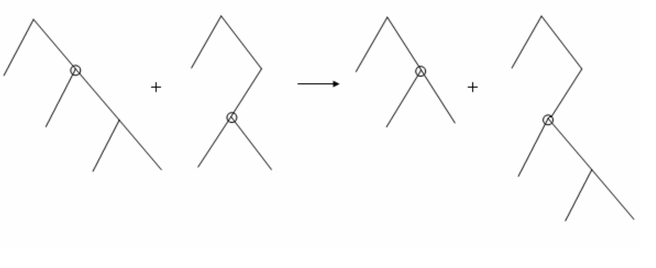

Fig. 2. Crossover operation

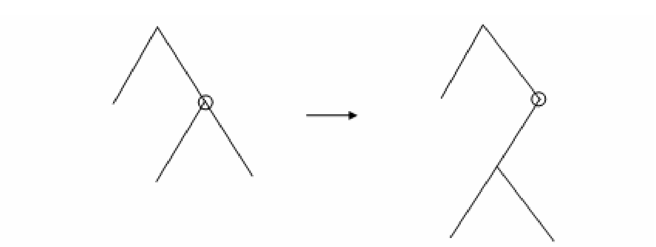

Fig. 3. Mutation operation

These processes, each of which occurs with a certain probability, are in operation from the first generation. The next generation is chosen from the new individuals thus created and from the previous generation according to a fitness function F, given as

$$F = \frac{1}{\sqrt{S+1}} - pN \, . \qquad \text{Equation 1}$$

Here, S is the score, which is the number of incorrect classifications; p is the node penalty, which is set at a fixed value; and N is the number of nodes in the tree. The purpose of the last term is to limit the size of the trees so that these do not grow out of proportions, in what is often referred to as code bloat, which makes the GP process take too long and the solution too complicated. The process continues until we have completed a certain number of generations. The final solution is the tree with the best score in the last generation obtained.

For the investigations in this paper, the parameters were set as follows:
population size = 100, number of generations = 300 (for smaller feature set) or 500 (for larger feature set) (see Sections IV and V), probability of mutation = 0.4, probability of crossover = 0.4, probability of replication = 0.2, and node penalty, p = 0.002.

IV. Feature Selection

Feature selection is often used in machine learning; it refers to the process where a subset of all the features extracted from the data is selected for use in a machine learning algorithm to make the problem more tractable. Five stand-alone feature selection algorithms (Kullback-Leibler divergence [KLD], Kolmogorov-Smirnov test [K-S test], Student's t test [t test], sequential forward selection [SFS] and sequential backward selection [SBS]) have been used in an initial step to narrow the pool of features for use with GP. To ensure a significant reduction from the number of original 22 features, nine (approximately 40%) of the top features were selected in all of the procedures. An important point to note here is that all of the methods have resulted in almost the same set of features; indeed the same top four features (features 5 to 8) result in each case (see Table 1).

Table 1
Results of selection of features by various procedures.

Feature-selection method or Statistical test	Ordered list of features*								
KLD	8	5	7	6	2	1	13	15	20
K-S test	8	6	5	7	2	4	1	17	10
t-test	8	5	6	7	16	2	9	1	17
SFS	8	5	7	6	2	1	16	9	18
SBS*	1	2	5	6	7	8	9	16	18

*For SBS, the features are given in numerical order rather than order of importance

Indeed the top four features (features 5 to 8) resulting from the tests are those that one would choose by simply reviewing the distributions of the feature values with respect to two classes in each feature. In Figure 4 are displayed values of the top four features (5 to 8) selected by all tests and the values of feature 13 that is selected by none of these tests. This plot demonstrates the efficacy of these tests.

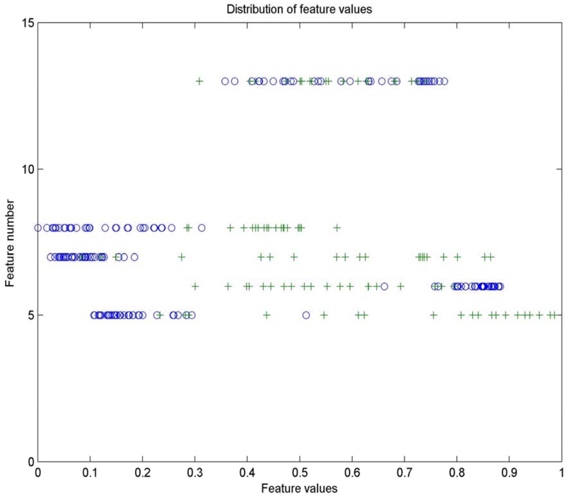

Fig. 4. Distribution of feature values for the 57 masses and tumors in the study. Malignant tumors are labeled by '+'; benign masses are labeled by 'o'.

In view of the results shown in Table 1, two new feature sets 'Y' and 'Z' were created: Y = {1, 2, 4, 5, 6, 7, 8, 9, 10, 16, 17}, containing eleven features, and Z = {2, 5, 6, 7, 8}, containing five features.

V. Experiments and Results

The original dataset available for the present study contains 57 ROIs – 37 related to benign masses and 20 of malignant tumors, and each ROI is represented by 22 feature values. The number of features is rather large compared with the number of data samples to be classified. In this case of a small, but very well categorized dataset, one cannot afford to set many cases aside for testing.

One method in common use in situations as above is the leave-one-out procedure [16], where the classifier is trained *n* times (*n* = number of data points), each time with a different data point left out for testing. The problem with this method is that leaving out one data point can drastically alter the results of estimation when one has a small amount of data to begin with; furthermore, there is a certain lack of confidence in the robustness in the measured performance.

In this paper, another approach is explored. A well-known statistical approach, bootstrap with resampling [17], is used to obtain statistical measures related to the performance of the classifier. Such a technique has been used in the analysis of breast masses [18]. We performed six experiments as shown in Table 2. Three experiments used the feature set Y and the other three experiments used the feature set Z. With each feature set (Y or Z) and for each condition (benign or malignant), the central *m*% of the data points in each feature space were selected. Using the sets of data points in each feature space selected as above, a new set of data points was created by performing either the union or the intersection of the sets of data points across all the features.

For example, Table 2 indicates that Experiment 1 was based on the feature set Y, and the dataset was obtained from the union of the central 5% of the samples in each feature space. Similarly, Experiment 6 was based on the feature set Z, and the dataset was obtained from the intersection of the central 80% of the samples in each feature space. In each experiment, 100 test sets were created by the well-known statistical procedure called sampling with replacement. Each test set comprised of five malignant tumors and five benign masses.

Table 2
Selection of data samples for classification experiments.

Experiment number	Feature set	Union or Intersection (U/I)	*m*
1	Y	U	5
2	Y	I	80
3	Y	I	90
4	Z	U	10
5	Z	I	70
6	Z	I	80

Table 3
Results of classification experiments with data sampling with replacement.

Experiment description*	Average Training performance	Average Test performance	Average overall performance
1_Y_U_5	99.9 %	90.1 %	98.2 %
2_Y_I_80	99.6 %	98.4 %	99.4 %
3_Y_I_90	100.0 %	99.5 %	99.9 %
4_Z_U_10	100.0 %	99.6 %	99.9 %
5_Z_I_70	99.9 %	100.0 %	99.9 %
6_Z_I_80	99.9 %	100.0 %	99.9 %

The notation for the description of each experiment is 'experiment number'_'feature set'_'union or intersection'_*m

There are two aspects to the following results - the performance of the classifier and the features that are found to be important. Using different feature sets, we have conducted six experiments. Results are presented in Table 3. It is found that the GP classifier correctly classifies the training data in over 99.5% of the cases and the test data in over 98% of the cases, except for the first experiment where the result is 90.1%. It should be remarked that the classification accuracies obtained are high and that the feature-selection step prior to the GP operation is helpful in that the results with one particular feature set appear to be more robust that those with other feature sets. This is a clear indication of the strength of GP with feature selection.

Having demonstrated the high performance of GP, we also explored the selection of important features. To determine which features are important for the purpose of classification using the GP classifier, the percentage of the number of times each feature was selected over all experiments for each feature set was calculated, and the most commonly selected feature combinations are also recorded for each feature set.

Table 4 shows the percentage of the number of times each feature was selected by the GP classifier using the feature set Z in Experiments 4, 5, and 6. It is clear that feature 8 (fractional concavity, F_{cc}) has been selected almost every single time in each of the experiments with the feature set Y or Z, and there is no other feature that is as discriminative as this one. Thus, we can make a firm statement that the shape factor fractional concavity is a strong measure for distinguishing benign masses from malignant tumors. This result agrees with the results obtained by Alto et al. [8]; however, the methods used in the two studies are different.

TABLE 4
THE PERCENTAGE OF SELECTION OF EACH FEATURE IN SET Z IN EXPERIMENTS 4, 5, AND 6.

Feature	2	5	6	7	8
Experiment 4	56	45	33	45	99
Experiment 5	54	56	28	53	99
Experiment 6	45	57	35	26	100
Average	51.7	52.7	32.0	41.3	99.3

VI. CONCLUSION

The GP classifier consistently performed well in discriminating between benign masses and malignant tumors using the shape features. The shape measure of fractional concavity (feature 8) was found to be the most important feature: it was selected almost every time by the GP classifier in all experiments. Accuracies obtained in classification of benign versus malignant using various combinations of the shape, edge-sharpness, and texture features are generally over 98%, which can be compared with results obtained by Alto et al. [8] using the same dataset but other classifiers such as K-nearest neighbors, Mahalanobis distance, linear discriminant analysis, and logistic regression.

Although the texture features were not favored by the statistical tests or feature-selection methods, two of the texture features in the feature set Y appear in the two most common feature combinations. This shows that the best set of two features is not necessarily made up of the two best features. Also, even if a classifier may not perform well with texture features on their own, it could perform well when using texture features combined with the shape features, a fact observed by Alto et al. [8]. It should be remarked that the estimation of shape factors requires accurate contours, which are not easy to obtain automatically. The contours of the masses employed to derive the features used in the present study were drawn manually by an expert radiologist specialized in mammography; regardless, questions arise regarding the dependence of the results upon the opinion of one expert, as well as the possibility of inter-observer and intra-observer differences. In addition, texture and edge-sharpness are important in radiological diagnosis, and there is a need for defining better features related to these aspects of breast masses in mammograms. It should be noted that the texture and edge-sharpness features used in this work are computed using bands of pixels around the given contour, which reduces the dependence of the features upon the accuracy of the contour to some extent.

ACKNOWLEDGMENT

This work was partly funded by the Medical Research Council, UK, through the Inter Disciplinary Bridging Awards (IDBA).

REFERENCES

[1] http://www.statistics.gov.uk/

[2] M J Yaffe, *Digital Mammography: IWDM 2000*; Madison, WI: Medical Physics Publishing, 2001.

[3] H–O Peitgen, *Digital Mammography: IWDM 2002*, Bremen, Germany: Springer-Verlag, 2003.

[4] R M Rangayyan, F J Ayres, and J E L Desautels, "Computer-aided diagnosis of breast cancer: Toward the detection of early and subtle signs," The First World Experts' Congress on Women's Health Medicine and Healthcare, World Academy of Biomedical Technologies, Paris, France, March 2005.

[5] R M Rangayyan, N M El-Faramawy, J E L Desautels, and O A Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. on Medical Imaging*, vol. 16, no. 6, pp. 799-810, 1997.

[6] B S Sahiner, H P Chan, N Petrick, R F Wagner, and L Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size", *Medical Physics*, vol. 27, no. 7, pp. 1509-1522, 2000.

[7] Alberta Cancer Board, "Screen Test: Alberta Program for the Early Detection of Breast Cancer," 2001/2003 Biennial Report, Edmonton, Alberta, Canada, 2004, http://www.cancerboard.ab.ca/screentest/

[8] H Alto, R M Rangayyan, and J E L Desautels, "Content-based retrieval and analysis of mammographic masses," *Journal of Electronic Imaging*, vol. 14, no. 2, Article No. 023016, pp 1 – 17, 2005.

[9] J R Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, USA, 1992.

[10] L Zhang, L B Jack, and A K Nandi, "Fault detection using genetic programming", *Mechanical Systems and Signal Processing*, vol. 19, pp. 271-289, 2005.

[11] H Guo, L B Jack, and A K Nandi, "Feature generation using genetic programming with application to fault classification," *IEEE Trans.* on *System, Man, and Cybernetics*, Part B, vol. 35, no. 1, pp. 89-99, 2005.

[12] J K Kishore, L M Patnaik, V Mani, and V K Agrawal, "Application of genetic programming for multicategory pattern classification", *IEEE Trans. on Evolutionary Computation*, vol. 4, no. 3, pp. 242-258, 2000.

[13] S Theodoridis and K Koutroumbas, "Pattern Recognition", Academic Press, City, State, USA, 2005.

[14] W H Press, B P Flannery, S A Teukolsky, and W T Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, UK, 1989.

[15] S Kullback and R A Leibler, "On information and sufficiency", Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79-86, 1951.

[16] R O Duda and P E Hart, "Pattern Classification and Scene Analysis", Wiley, New York, NY, 1973.

[17] B Efron and R J Tibshirani, An Introduction to the Bootstrap, CRC Press LLC, Boca Raton, FL, 1998.

[18] Y Liu, M R Smith, and R M Rangayyan, "The application of Efron's bootstrap methods in validating feature classification using artificial neural networks for the analysis of mammographic masses", 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society , San Francisco, CA: IEEE, pp. 1553-1556, 2004.