Adv Syst Sci Appl 2018; **4**; 12-19
Published online at http://ijassa.ipu.ru/

Severity of Breast Masses Prediction in Mammograms Based on Optimized Naive Bayes Diagnostic System

Abeer S. Desuky

Faculty of Science, Al-Azhar University, Cairo, Egypt

E-mail: abeerdesuky@azhar.edu.eg

Abstract. Mammography is the most effective tool for breast masses screening. It is a special CT scan technique used only to detect breast tumors early and accurately. Detecting tumors in its early stage has improved the survival rate for breast cancer patients. Computer aided diagnostic systems help the physicians to detect breast cells abnormalities earlier than other traditional procedures. In this paper, an improved Naive Bayes classifier based on Chicken swarm optimization algorithm (CSO-NBC) is analyzed on mammographic mass dataset. The main aim of this research is to increase physician's ability to determine the severity of a mammographic mass lesion from the BI-RADS features and the patient's age using the bio inspired chicken swarm optimization (CSO) algorithm for Naive Bayes classifier (NBC) improvement. The dataset is preprocessed and divided to train the CSO-NBC system and test it by 5-folds cross validation technique. The performance of our proposed classification system is compared with papers' results of other researchers to show the efficiency of our system in predicting severity of breast tumors with the highest accuracy.

***Keywords*:** Breast Masses Mammography, Naive Bayes, Chicken Swarm Optimization, Computer Aided Diagnosis.

1. INTRODUCTION

Cancer is one of the top leading causes of death. It is caused by uncontrolled growth of cells which invade and spread around the body, often resulting in death. Cancer is the second leading cause of death globally, and was responsible for 8.8 million deaths in 2015 [1]. Globally, nearly 1 in 6 deaths is due to cancer according to the WHO (World Health Organization). Per the US National Cancer Institute, 60 percent of the world's new cancer cases happen in Asia, Africa, Central and South America, and 70 percent of global cancer deaths occur in those same regions as well. Moreover, breast cancer death-rate was 571 000 equivalently 6.5 percent of all deaths in 2015 [2]. Because of this fact, early detection and true diagnosis is an important issue and plays a key role to reduce mortality of this disease.

Mammography is an efficient imaging appliance for early breast cells abnormality detection. Mammography exists in two types: first type, screening mammography which is breast X ray used to test the changes in breast area for early detection of breast tumors in women with no symptoms of cancer. It can also detect tiny calcium deposits (micro calcification) that is one of cancer manifestations. The second type, Diagnostic mammography which is a breast X ray to check the signs of breast cancer after detecting a mass or other symptom. Symptoms include breast size/shape change, skin thickening, pain or nipple discharge [3, 4].

Significant improvements can be made in the lives of breast cancer patients by detecting cancer early and avoiding delays in care, computer aided diagnosis (CAD) can help the physicians to do this task faster and more accurate than traditional procedures. During the last decade with development of machine learning approaches in diagnostic systems, breast cancer detection has improved. The aim of using machine learning approaches is to minimize mistakes that may occur by specialists in diagnosis [5].

Different methods have been proposed to detect and classify masses in mammogram images. Ramani and Vanitha [4] used weighted histogram algorithms to select features from mammogram data then classified the selected features using random forest, naive Bayes and ANN algorithms. Another method proposed in [6] based on missing values imputation and three models SVM with polynomial kernel, ANN with pruning parameters and DT with Chi-squared interaction detection were derived for prediction.

Bio-inspired swarm algorithms can be used to enhance effectiveness of CAD systems. This paper presents the bio inspired chicken swarm optimization (CSO) algorithm to improve naive Bayesian classifier (NBC) for the severity of breast masses prediction in mammograms.

2. THE PROPOSED METHOD

2.1. Chicken swarm optimization

Chicken swarm optimization (CSO) algorithm proposed by Meng et al. [7], is a bio-inspired swarm algorithm which mimics the behavior of chicken swarms in nature. Each chicken swarm contains several groups of one rooster with two or more hens and many chicks move in a hierarchy groups and searching for food which has an effect on the swarm movements.

Dividing the chicken swarm into groups and identifying the chickens into roosters (R), hens (H) and chicks (C) mainly depends on the fitness values of the chickens. The chickens with best fitness values would be the roosters.

Chicks would be the chickens with worst fitness values while the others would be the hens (mothers). These statuses are updated according to the change in the fitness values only every several (G) time steps. There are three basic movements in the swarm as proposed in [7]:

First Movement is Rooster's movement which has the better fitness values since it can search for food in a wider range than the rest of chicks with less fitness values, the movement of the rooster is defined in Eq. 1 and Eq. 2:

$$x_{i,j}^{t+1} = x_{i,j}^{t} * \left(1 + randn(0, \sigma^2)\right) \tag{1}$$

Where *x* is the selected rooster, $randn(0, \sigma^2)$ is a Gaussian distribution with mean 0 and standard deviation σ^2, and *f* is the fitness value of the corresponding rooster *x*.

$$\sigma^2 = \begin{cases} 1 & f_i \leq f_k \\ \exp\left(\frac{f_k - f_i}{|f_i| + \varepsilon}\right) & k \in [1, N], k \neq i \end{cases} \tag{2}$$

Where *k* is rooster index, *i* is the related position and ε is the smallest integer number in the computer used to avoid zero-division-error.

Second Movement is Hen's; Hens follow the roosters to search for food and their movement is defined in Eq. 3 – Eq. 5:

$$x_{i,j}^{t+1} = x_{i,j}^{t} + W1 * rand * \left(x_{r1,j}^{t} - x_{i,j}^{t}\right) + W2 * rand * \left(x_{r2,j}^{t} - x_{i,j}^{t}\right) \tag{3}$$

$$W1 = \exp\left(\frac{f_i - f_{r1}}{abs(f_i) - \varepsilon}\right) \tag{4}$$

$$W2 = \exp(f_{r2} - f_i) \tag{5}$$

Where $r1$ and $r2$ are the indices of the rooster and the randomly chosen chicken (rooster or hen), r1 ≠ r2, and *rand* is a uniform random value between [0, 1].

Third Movement is the movements of the chicks which follow their mothers, defined in Eq. 6:

$$x_{i,j}^{t+1} = x_{i,j}^{t} + L * \left(x_{m,j}^{t} - x_{i,j}^{t}\right) \tag{6}$$

Where $x_{m,j}$ is the position of the chick's mother, and L is a randomly chosen parameter between 0 and 2.

2.2. Naive Bayesian Classifier

The Bayesian classifier represents a statistical classification algorithm as well as a supervised learning method for classification based on Bayes theorem [8]. Assuming a probabilistic model, in Bayes theorem, posterior probability of a data sample x with unknown class label is calculated as in Eq. 7.

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)} \tag{7}$$

Where each data sample is represented by an n-dimensional feature vector, $x = (x_1, x_2 .... x_n)$ and h is some hypothesis, such that the data sample x belongs to a specified class label $l_1, l_2 .... l_m$. The NB classifier will predict that the given data sample x belongs to the class label l_i that have the higher posterior probability, i.e. for $p\ (l_i \mid x)$ is the maximum [9].

$$p(l_i|x) > p\left(l_j\middle|x\right) \text{ for } 1 \leq \text{j} \leq \text{m} \tag{8}$$

2.3. The dataset

Researchers proposed several (CAD) computer aided diagnostic systems in the last few years. These systems help doctors in deciding to perform short term follow-up examination or proceed a breast biopsy on a suspicious lesion seen in a mammogram instead.

Table 1. Mammographic Mass Dataset Features

Feature	Type	Values and labels		Missing values
BI-RADS	Ordinal (non-predictive)	0	Assessment incomplete	2
		1	Negative (non-predictive)	
		2	Benign findings	
		3	Probably benign	
		4	Suspicious abnormality	
		5	Highly suggestive of malignancy	
Age	Integer	18 - 96 (Patient's age in years)		5
Shape	nominal	1	Round	31
		2	Oval	
		3	Lobular	
		4	Irregular	
Margin	nominal	1	Circumscribed	48
		2	Microlobulated	
		3	Obscured	
		4	Ill-defined	
		5	Spiculated	
Density	Ordinal	1	High	76
		2	Iso	
		3	Low	

		4	Fat-containing	
severity	Binominal	0	Benign	0
	(goal field)	1	Malignant	

BI-RADS (Breast Imaging Reporting and Data System) is developed to standardize nomenclature between radiologist and doctors. BI-RADS assessment is done based on shape, age, margin density. BI-RADS features and the patient's age in mammographic data set can be used to predict the severity of a mammographic mass lesion.

This data set contains the patient's age, a BI-RADS assessment, and three BI-RADS features together with the class feature (severity) for 445 malignant and 516 benign masses that have been collected and identified on full field digital mammograms at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. Table 1 shows that each instance has BI-RADS features, age, shape, margin, density and severity features. Mammographic data set includes 162 missing values [6, 10].

2.4. Naive Bayes Based on Chicken Swarm Optimization

Traditional NBC is a statistical classifier with mutual independency among the features [4]. In the medical domain, all the symptoms do not contribute equally in diagnosing a specific disease. For example, in heart disease domain, the BMI feature is having less impact than the prior-stroke feature in predicting the probability of being heart patient [9]. In our proposed work called CSO-NBC we assign weights to each feature according to their efficiency in prediction. we use CSO technique to learn the features weight in NBC. No information or assumptions are needed about the weight in NBC, the mechanism of CSO can help us get automatically the optimal weights.

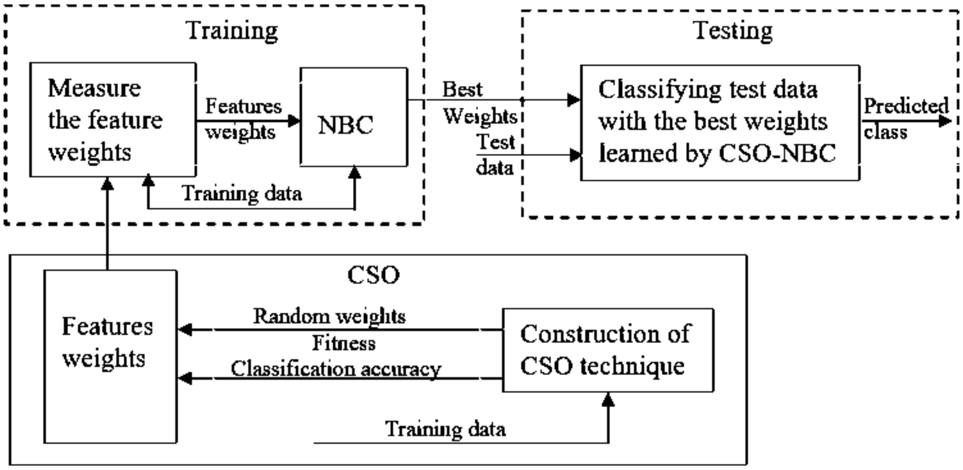

Fig. 1 CSO-NB Classification technique

We use the weight to increase the classification performance of the NBC. A high weight will be assigned to features that have more impact in prediction and low weight are assigned to features that have less impact. So, our main aim is to get optimum weight configuration that gives the highest accuracy for CSO-NBC. The major steps of the working technique are shown in Figure 1 and described as follows:

First step: the mammographic mass data is preprocessed divided into training and testing sets then, the NBC applied on it firstly to measure the NB classification performance.

Second step: an initial generation for weight individuals (chickens) is formed in a random manner. Each chicken swarm is a vector of all weight variables ranging from -10 to 10. For the weight individuals, the generation size N should be determined first.

Third step: The total generation cost of each chicken is calculated.

Fourth step: Generate new rooster around the global best swarm by searching for the best fitness index and the worst fitness index of chicken swarm in same iteration and store their indexes; then, CSO-NBC replaces the worst index with the best index.

Fifth step: Third and fourth steps are repeated till get the best individual weight which is the one getting the highest classification accuracy.

Last step: the test instances can be classified under the learned weight by CSO-NBC and the performance measures can be determined.

3. EXPERIMENTAL RESULTS

The effectiveness of any diagnostic system is evaluated by its ability to give the maximum accurate classification result [11]. Per the real nature for any given case and the prediction from diagnostic system, there are only four possible outcomes, true negatives (TN) as well as true positives (TP) correspond to a correct diagnosis; that is, cases are successfully labeled as uninfected and infected patients, respectively; false positives (FP) refer to uninfected patient being classified as infected; false negatives (FN) are infected patients incorrectly classified as uninfected. The most popular performance metrics is accuracy which can be calculated as:

$$Accuracy = \frac{\mathrm{TN+TP}}{\mathrm{TN+TP+FN+FP}} \quad (9)$$

To evaluate our system (CSO-NBC), besides the classical Accuracy, the standard metrics of Sensitivity, G-mean (geometric mean), and F-measure have been used.

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (10)$$

$$\mathrm{G-mean} = \sqrt{\mathrm{TPR} \cdot \mathrm{TNR}} \quad (11)$$

$$\mathrm{F-measure} = \frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}} \quad (12)$$

A diagnosis system should have a high accuracy, Sensitivity, G-mean and F-measure [4,12].

The experiments were implemented on a computer with Intel Core™ i7 processor and 6GB RAM running Microsoft Windows 7 Professional and the algorithm is coded in MATLAB 15.

First, we replaced all the missing features values - substitute it with the average - in the mammographic mass dataset then, applied Information Gain (IG) feature selection technique [13], which is used to evaluate the importance of a feature by measuring the Information Gain (entropy) with regard to the class - to remove useless features and improve the classification performance. The instances are randomly divided into training and testing sets in a 5-folds cross validation manner. The used parameters in evolution were: 100 for population size, 1000 generations (iterations) and time steps (G=10).

Tables 2, 3, 4 and 5 show the performance evaluation: classification accuracy, sensitivity, G-mean and F-measure respectively - of the NB classifier and the proposed technique using the full mammographic mass dataset features and the subset of selected features after applying Information Gain technique and removing “Density” feature.

The results show a perceivable improvement in classification performance for CSO-NBC and show that with IG feature subset selection our proposed technique enhanced the classification performance against the enhancement with NB classifier.

Table 2 classification accuracy

	NB	**CSO-NBC**
Full data	78.87	83.88
Reduced data	85.42	87.20

Table 3 Sensitivity

	NB	**CSO-NBC**
Full data	70.16	89.52
Reduced data	87.57	82.94

Table 4 G-mean

	NB	**CSO-NBC**
Full data	78.81	82.63
Reduced data	85.11	87.37

Table 5 F-measure

	NB	**CSO-NBC**
Full data	77.93	85.80
Reduced data	86.48	87.36

Table 6 Comparison of algorithms Applied on Mammographic Mass Data

Method	**Accuracy**
Proposed CSO-NBC	87.20
NB [5]	82.49
SVM [6]	81.25
Bagging SVM-SMO [9]	82.00
Bagging DT [9]	83.40

The experiments also showed - Table 6 -that the proposed algorithm CSO-NBC outperforms algorithms proposed by other researchers applied on the mammographic mass data.

4. CONCLUSION

Mammography is used to help in the early detection (diagnosis) of breast diseases. The Computer Aided Diagnostic systems can help physicians in detection (diagnosis) of abnormalities early than traditional procedures. Many algorithms and diagnostic systems have been developed recently. The objective is not to replace medical researchers and professionals, but to increase their abilities to take decisions about the disease. This paper proposed a new automated mammogram classification system CSO-NBC based on the improved Naive Bayes classifier using Chicken swarm optimization algorithm. The mammographic mass data is preprocessed by replacing all the missing features values and applying Information Gain (IG) feature selection technique to remove the useless features. The dataset then divided to training and testing sets using 5-folds cross validation technique. The experiments showed the efficiency of our proposed classification system in predicting severity of breast tumors since it performs better than the traditional NB classifier also, better than algorithms proposed by other researchers for the same mammographic mass data.

REFERENCES

[1]World Health Organization (2018, May) Cancer [Online]. Available: http://www.who.int/cancer/en/

[2] World Atlas (2018, May) Explore The World [Online]. Available: http://www.worldatlas.com

[3] Sickles, E. A., Wolverton, D. E., & Dee, K. E. (2002). Performance Parameters for Screening and Diagnostic Mammography: Specialist and General Radiologists. *Radiology*, 224(3), 861-869.

[4] Ramani, R. & Suthanthira Vanitha, N. (2014). Computer Aided Detection of Tumors in Mammograms, *Int.J. Image, Graphics and Signal Processing*, 4, 54-59.

[5] Güzel, C., Kaya, M. & Yıldız, O. (2013). Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation, *AWERProcedia Information Technology & Computer Science*, 4, 401-407, [Online].
Available: www.awer-center.org/pitcs

[6] Mokhtar, S. A., Elsayad, A. M. (2013). Predicting the Severity of Breast Masses with Data Mining Methods, *International Journal of Computer Science Issues*, 10(2).

[7] Meng, X., Liu, Y., Gao, X. & Zhang, H. (2014). A New Bio-inspired Algorithm: Chicken Swarm Optimization. In: Tan Y., Shi Y., Coello C.A.C. (Eds.) *Advances in Swarm Intelligence. ICSI 2014. Lecture Notes in Computer Science*, vol 8794. Springer, Cham, https://doi.org/10.1007/978-3-319-11857-4_10

[8] Ozer, P. (2008). Data Mining Algorithms for Classification, *B.Sc Thesis*, Redbound University Nijimegan.

[9] Jeyarani, D. S., Anushya, G., Raja, R. & Pethalakshmi, A. (2013). A Comparative Study of Decision Tree and Naive Bayesian Classifiers on Medical Datasets, *International Journal of Computer Applications (0975 – 8887), International Conference on Computing and information Technology* (*IC2IT*-2013), 5-7.

[10] Luo, S.-T., Cheng, B.-W. (2012). Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods, *Journal of Medical Systems*, 36(2), 569-77. https://doi.org/10.1007/s10916-010-9518-8.

[11] El Bakrawy, L. M. & Desuky, A. S. (2015). A hybrid classification algorithm and its application on four real-world data sets, *International Journal of Computer Science and Information Security (IJCSIS)*, 13(10): 93-97.

[12] Han, J. & Kamber, M. (2001). *Data Mining Concepts & Techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[13] Kumar, C. S. & Sree, R. J. (2014). Application of ranking based attribute selection filters to perform automated evaluation of descriptive answers through sequential minimal optimization models, *ICTACT Journal on Soft Computing*, 5(1), 860-868.