

MULTI-CLASS FEATURE SELECTION ALGORITHM (MCFSA) FOR BREAST CANCER DETECTION

¹P. Sherubha, ²M. Banu Chitra, ³B. Narmadha,

^{1,2} Assistant Professor, Department of Information Technology,
Karpagam College of Engineering, Coimbatore.

³Knowledge Institute of technology, Salem,

¹sherubha0106@gmail.com, ²banuchitracs@gmail.com, ³bnarmadha29@gmail.com

Abstract: In this paper, a new feature selection algorithm has been proposed termed Multi-Class Feature Selection Algorithm (MCFSA) for effective feature selection to be used in a new cancer detection system. In addition, an adaptive Neuro-Fuzzy k-NN Classification algorithm was utilized for testing the performance of the feature selection algorithm in terms of classification accuracy. To test the efficacy of the proposed medical expert system, the Mini-MIAS Dataset (M-MIAS) has been used. The obtained classification accuracy is a very promising result compared to the existing works in this area based on the reports from the results for the same data set.

Keywords: Adaptive Neuro-fuzzy, Feature selection, Classification, Multi-class, Nearest neighbor.

1. Introduction

Breast cancers are one of the most usually taking place cancers in girls. Early analysis of breast most cancers is crucial and critical in lowering mortality charge and improving the patient's quality of life. An early analysis of this disorder has more significance and drastically improves the diagnosis and results in extra effective remedy for the affected person. Expert machine is a pc system that makes use of artificial intelligence to emulate the choice-making ability of a human professional and to resolve issues inside a specialized domain that ordinarily requires human knowledge. In order to attain the feats of apparent intelligence [5]-[6], an expert machine is based on additives referred to as understanding base and an inference engine. An understanding base is an organized as a collection of facts approximately the device's domain. An inference engine interprets and evaluates the facts in the expertise base as a way to provide a powerful solution. Typical tasks for expert systems involve category, prognosis, monitoring, and layout, scheduling, and making plans for specialized endeavors [7].

1.1 Mini-MIAS Dataset

The Mammography Image Analysis Society (MIAS) is an agency of UK studies groups that have produced a digital mammography database. The pictures are in gray scale document format. The original MIAS Database (digitized at 50 micron pixel part) has been reduced to two hundred micron pixel edge and clipped/padded so that each image is 1024 pixels x 1024 pixels called the mini-MIAS database. This effort have used mini-MIAS database because it carries entire statistics approximately abnormalities of each mammographic images.

2. Literature survey

Hanung Adi Nugroho et al. anticipated unique accuracies, specificities, and sensitivities are obtained as effects from the type shape of breast ultrasound images the usage of SVM and MLP. At the Zernike second, the very best accuracy received through the usage of SVM classifier is eighty four.80%, while the use of the MLP classifier is 87.90%. At invariant moment, accuracy received the usage of SVM classifier is 69.Sixty nine %, while the use of the MLP classifier is 78.70%. On the opposite aspect, whilst Zernike moment and invariant moment are mixed, the type result achieves ninety 3.9% for accuracy, 91.70% for specification, and 100% for sensitivity.

Uma Ojha et al. anticipated the prediction version to classify recurrent or non-recurrent instances of breast cancer is a research that is statistical in nature. Still these images can be related to bio clinical evidences. In this paper, WPBC dataset is used for finding an efficient predictor set of rules to predict the recurring or non-ordinary nature of ailment. This would possibly assist Oncologists to differentiate a terrific diagnosis (non-recurrent) from a terrible one (recurrent) and might treat the sufferers more successfully. Eight popular information mining strategies have been used, four from clustering algorithms (K-means, EM, PAM and Fuzzy c-method) and four from classification

algorithms (SVM, C5.Zero, KNN and Naive Bayes). The outcomes of these algorithms are genuinely outlined in this paper with necessary consequences. The category algorithms, C5.0 and SVM have proven 81% accuracy in classifying the recurrence of the disorder.

Priya Darshini Velusamy et al supplied the advanced adaptive fuzzy C way algorithm together with modified area growing set of rules which gives the best result for the images which has more facts and relatively better than k-manner and ordinary Fuzzy C approach set of rules because in k-means the facts point in particular belong to 1 cluster center but in advanced AFCM the records factor has assigned club that is exact to each cluster middle in order that the facts factor will belongs to a couple of cluster center. Also we've got calculated the area, perimeter and diameter of the tumor as a way to discover the exact portion and the edges of the tumor.

V. Vishrutha et al. Proposed approach the usage of mixture of wavelet functions and texture features improves the accuracy of CAD system for detection of early level breast cancer. We effectively proposed and implemented a way to locate micro calcifications and circumscribed masses which imply early stage breast most cancers and classify them as benign and malignant the usage of SVM classifier with an accuracy of ninety two%. SVM classifier affords greater accurate and particular outcomes. This research may be further advanced to work with distinct wavelet based totally functions and with greater texture capabilities that may result in better efficiencies.

Hamid R. Tizhoosh et al. supplied ROI estimation in ultrasound B-mode pictures of regionally superior breast cancer. It turned into validated that the usage of international and nearby Radon barcodes ought to offer a quick semi-automatic difficult localization of tumors in modalities like ultrasound in which noise and the inherent vagueness of tumor borders make contour extraction extraordinarily hard.

3. Proposed work

This work discusses about the proposed feature selection algorithm called multi-class feature selection algorithm and an adaptive Neuro-Fuzzy k-NN classifier.

In this proposed feature selection algorithm, first calculate the Data Acquisition Ratio (DAR) value for all features by using the equations 1, 2 & 3.

$$DAR(X_i) = \frac{I_D - I_E}{I_D + I_E} \times 100 \quad (1)$$

where,

$$I_D = - \left(\frac{freq(C_j, D)}{|D|} \right) \log_2 \left(\frac{freq(C_j, D)}{|D|} \right) \quad (2)$$

$$I_T = \left(\frac{|T_i|}{|T|} \right) \times I(T_i) \quad (3)$$

Where,

$D \rightarrow$ the total dataset,

$C_j \rightarrow$ particular feature of the cluster (Subset),

$T \rightarrow$ total training dataset considered for DAR calculation,

$T_i \rightarrow$ features of the training dataset.

$A_i \rightarrow$ attribute (feature) of the particular record.

3.1 Feature Selection

The proposed weighted subset feature selection algorithm is proposed with the help of efficient optimization technique called particle swarm optimization algorithm and effective feature subset based classifier called Enhanced Multiclass decision tree algorithm for validating the classification accuracy of selected feature subset.

In PSO, an individual in the swarm, called a particle, represents a potential solution. Each particle has a fitness value and a velocity, and it learns the experiences of the swarm to search for the global optima as in figure 1.

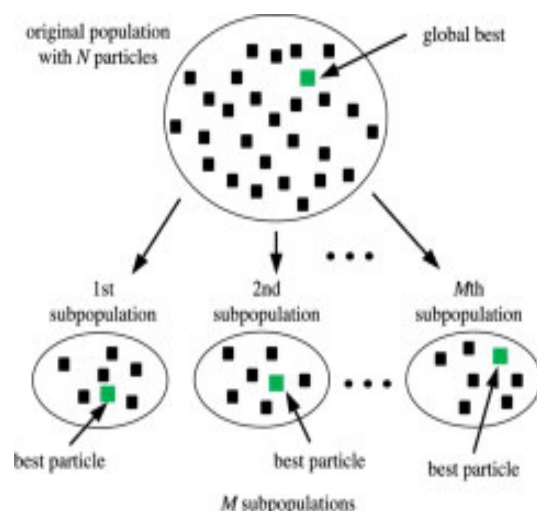


Figure 1. Representation particle swarm optimization

3.2 Enhanced multiclass Decision tree

Enhanced Multiclass Decision Tree (EMDT) has been used for evaluating the selected feature subsets during the process of feature selection as in figure 2. In this work, this classification algorithm helps to calculate the detection accuracy based on many aspects of the problem [1]. It is capable of solving the multiclass problem efficiently.

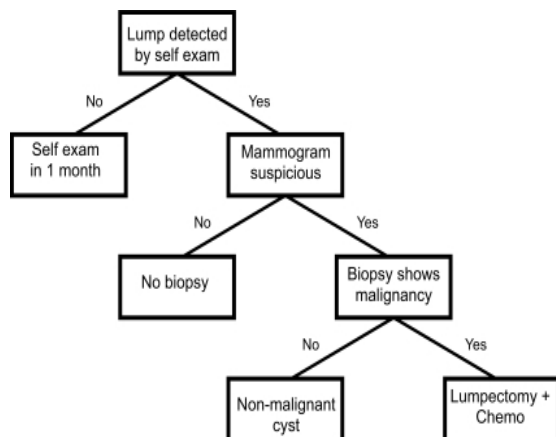


Figure 2. Representation tree separation using decision tree

The multi-class feature selection algorithm (MCFSA). In this paper, proposes a new feature selection algorithm called Multi-Class Feature Selection Algorithm (MCFSA) with the help of DAR, Particle Swarm Optimization Algorithm and EMDT. Here, the proposed algorithm chooses a set of optimal features as input to the Particle Swarm optimization algorithm for selecting the valuable features from the full set of features. These selected features are evaluated by using EMDT. If the selected feature subset provides better classification accuracy then the process is stopped otherwise it is repeated until it provides classification accuracy above a prescribed level.

The anticipated feature selection algorithm works as:

Input: F-features set, DAR: Data Acquisition Ratio, C: Classifier, T: Threshold

Output: Optimal features

Algorithm:

Step 1: F_i , read the features

Step 2: Calculate the Data Acquisition Ratio for all features using the equations (1-3).

Step 3: Rank features based on DAR feature value.

Step 4: Initiate an empty set $S = \{\}$, $ax=0$;

Step 5: Initiate the value of ac in to yx .

Step 6: Read subsequent feature and allocate temporary variable f .

Step 7: Add the feature to the latest feature Subset.

Step 8: Eliminate the feature from unique feature set.

Step 9: Compute new feature subset accuracy with EMDT [3] and amass into ax .

Step 10: If $((ax-yx) < T$ OR $ax < yx$) then Go to Step 10.

Step 11: or else re-rank the chosen feature subset using PSO [2]

Step 12: Replicate step 5 to Step 9

Step 13: Record optimal features.

The anticipated algorithm assists to recognize the optimal features [2] which are supportive for assembling effectual decisions over the medical dataset [3]. The anticipated algorithm offers better prediction accuracy over the medical dataset owing to the utilization of multi-classed features.

3.3 Classification

In this investigation, classification is performed using an adaptive Neuro-Fuzzy classification algorithm over the selected features. Furthermore, this procedure has been tested using MIAS dataset. The anticipated feature selection algorithm and the classifier utilized in this effort, it is experiential that the anticipated feature selection procedure used with the classifier has abridged the classification time and enlarged the classification precision.

The classification algorithms are utilized to classify the data. The rules are cast off to confirm the grouping of the information. Certain classification algorithms such as Multi-class Decision tree make use of distance measurement procedure to locate the distance among the data value and believe this value for assembling earlier decisions.

After considering the preliminary decision, the last decision will be considered using the rules framed in accordance to the crisis by the algorithm. These works have utilized MCDT for creating effectual choice over the MIAS dataset. For enhancing the classification precision of the prevailing classification algorithm [8], a novel feature selection algorithm has been proposed called Multi-Class feature selection procedure.

The anticipated feature selection algorithm assists to progress the classification precision by means of optimal sum of features [4]. The anticipated feature selection procedure decreases the classification time.

3.4 Breast tumor classification

This work recommends MATLAB program to develop the categorization and recognition of breast tumor over mammograms, the classification phases are:

1. Verify whether the image comprises of tumor or not.
2. If no tumor is obtained then it is a normal image.
3. If tumor exists then it is an abnormal image.
4. Image segmentation
 - Transfer RGB to gray.
 - Relate the Weiner filtering to eliminate the noise in the image.
 - Identify the boundaries of image
5. Classifying the tumor in accordance to the tumor shape, and edge sharpness:
 - Detach the tumor area from the image
 - If tumor area is oval or round or lobulated then it is Benign tumor.
 - If tumor area is stellate or irregular or nodular then Malignant tumor.

The anticipated scheme is carried out in MATLAB program. The subsequent figure projected certain output of mammogram images and the MATLAB program experimentation to classify them as: normal, benign, and malignant.

4. Simulation Results and Discussion

The Mini Mammographic (MIAS) breast cancer data set has been used which have been taken from the UCI Machine Learning repository to assess the effectiveness of the anticipated classification model. Hence, it is estimated that the anticipated feature selection procedure using MIAS dataset. This work have considered three evaluation criteria's as in table I and in figure 3 like sensitivity, specificity, and accuracy for obtaining the classification precision.

Table 1. Comparison table for sensitivity, specificity and accuracy of diverse classifiers

Classifier	Sensitivity	Specificity	Accuracy
ANF k-NN	97.00%	3.20%	98.50%
NFS	95.70%	4.60%	95.80%
MLP	84.20%	13.20%	84.20%
SVM	87.50%	11.70%	87.40%

The experiments have been accomplished using MATLAB. These researches have measured the MIAS dataset for estimating the anticipated algorithm. The typical dataset is separated into two sets (70% and 30%), former for training and latter for testing. Experiments have been performed for computing the anticipated scheme with diverse datasets with features. Figure 3 illustrates the recital examination of the anticipated system with features and chosen features.

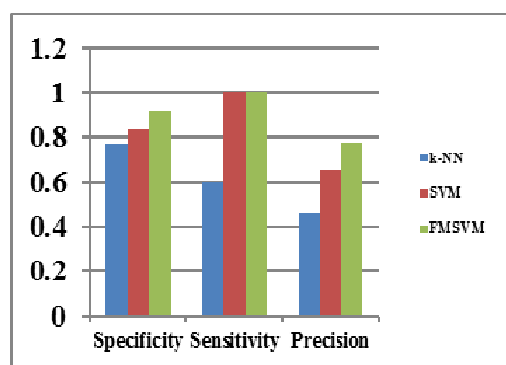


Figure 3. Computation of the specificity, sensitivity and precision using diverse classifiers

5. Conclusion and Future Enhancement

In this work, a novel Multi-Class feature selection algorithm for effectual feature selection to accomplish the improved classification precision over the available dataset is anticipated. By initiating the novel feature selection algorithm, diverse classification algorithms were experienced to discover the classification accuracy. The beginning of multi-class feature selection is extremely functional to recognize the most significant features for improved classification accuracy. Lastly, the finest classifier identified as adaptive Neuro-Fuzzy Classification algorithm is applied in this work for identifying breast cancer. The major involvement of this effort is the initiation of a novel algorithm for feature class which performs the feature ranking professionally for attaining improved classification precision [9]-[10]. The classification accuracy distinction among the anticipated system and the prevailing systems is approximately 1%.

References

- [1] Sindhu SS, Geetha S, Kannan A. Decision tree based light weight intrusion detection using a wrapper approach. *Expert Syst Appl* 2012; 39: 129-141.
- [2] Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, San Jose, CA, 1993; 1905: 861-870.
- [3] Mangasarian OL, Street WN, Wolberg WH. Breast cancer diagnosis and prognosis via linear programming. *Operat Res* 1995; 43: 570-577.
- [4] Buchsbaum RJ, Oh SY. Breast Cancer-Associated Fibroblasts: Where We Are and Where We Need to Go. *Cancers* 2016; 8: 1-19.
- [5] Mani NL, Schalper KA, Hatzis C, Saglam O, Tavassoli F, Butler M, Chagpar AB, Pusztai L, Rimm DL. Quantitative assessment of the spatial heterogeneity of tumor-infiltrating lymphocytes in breast cancer. *Breast Cancer Res* 2016; 18: 1-10.
- [6] Ganapathy S, Sethukkarasi R, Yogesh R, Vijayakumar P, Kannan A. An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization. *Sadhana* 2014; 39: 283-302.
- [7] Di Leo A, Curigliano G, Dieras V, Malorni L, Sotiriou C, Swanton C, Thompson A, Tutt A, Piccart M. New approaches for improving outcomes in breast cancer in Europe. *Breast* 2015; 24: 321-330.
- [8] Ganapathy S, Yogesh P, Kannan A. An Intelligent Intrusion Detection System for Mobile Ad-

Hoc Networks Using Classification Techniques. Comput Commun Informa Syst 2011; 148: 117-122.

[9] El-Khatib K. Impact of Feature Reduction on the Efficiency of Wireless Intrusion Detection Systems. IEEE Transact Parallel Distribut Syst 2010; 21: 1143-1149.

[10] Ghaemi M, Feizi-Derakhshi MR. Forest Optimization Algorithm. Exp Syst Appl 2014; 41: 6676-6687.

[11] P.R.Ganesh, V. Chandiraprakash," A Data Integral Multiple Cloud Storage Using Erasure Codes as Fault Tolerant System", International Innovative Research Journal of Engineering and Technology, Vol 01, No 01, pp.01-05,2016.

[12] S.V.Manikanthan and V.Rama "Optimal Performance of Key Pre distribution Protocol In Wireless Sensor Networks" International Innovative Research Journal of Engineering and Technology ,ISSN NO: 2456-1983,Vol-2,Issue –Special –March 2017.

[13] Rajesh, M., and J. M. Gnanasekar. "Constructing Well-Organized Wireless Sensor Networks with Low-Level Identification." World Engineering & Applied Sciences Journal 7.1 (2016).

[14] T.Padmapriya, Ms. N. Dhivya, Ms U. Udhayamathi, "Minimizing Communication Cost In Wireless Sensor Networks To Avoid Packet Retransmission", International Innovative Research Journal of Engineering and Technology, Vol. 2, Special Issue, pp. 38-42.

