Comparison of a Classifier Performance Testing Methods: Support Vector Machine Classifier on Mammogram Images Classification

Sura Jasim Mohammed
College of Sciences, Al Mustansiriyah university,Baghdad, Iraq
thekra.abbas@yahoo.com

Thekra Hayder Abbas
College of Sciences, Al Mustansiriyah university, Baghdad,Iraq
suragS.dcc@gmail.com

Received Nov. 12, 2018. Accepted for publication Dec. 23, 2018

DOI : *http://dx.doi.org/10.31642/JoKMC/2018/060102*

Abstract— *This paper compares between testing performance methods of classifier algorithm on a standard database of mammogram images. Mammographic interchange society dataset (MIAS) is used in this work. For classifying these images tumors a multiclass support vector machine (SVM) classifier is used. Evaluating this classifier accuracy for classifying the mammogram tumors into the malignant, benign or normal case is done using two evaluating classifier methods that are a hold-out method and one of the cross-validation methods. Then selecting the better test method depending on the obtained classifier accuracy and the running time consumed with each method. The classifier accuracy, training time and the classification time are considered for comparison purpose.*

Keywords: mammogram; multiclass SVM;classifier test; cross-validation; and hold-out.

I. INTRODUCTION

Advanced strides have been reached in the medical field last decade. Particularly wonderful results have been attained in the image classification area. here are in building a computer-aid diagnosis (CAD) system for breast cancer diagnosing image. This classification results from a combination of a sequence of applied techniques in each step coming before the classification step. In this paper, the images used in this system has been subjected to a set of operations and techniques including preprocessing using histogram equalization image enhancement, segmentation using Otsu's threshold, features extraction by applying wavelet discrete transform, these extracted features are reduced its dimensionality using principal components analysis (PCA) approach, final step it is the classification or as known as pattern recognition to recognize whether the suspected tumor is malignant, Benign or normal, this recognition is done after passing the desired image's features through multiclass SVM classifier. Early detection of breast cancer is often identified via masses and microcalcifications type prediction. Prediction of masses type (classification) using SVM was proposed and implemented with accuracy rate 94.79% % (Eltoukhy et al., 2012) [1].

Each classification system includes classification, test and training processes. Once the classifier has applied and chosen for a classification algorithm's system, its true error rate (accuracy) should be estimated. Classifier accuracy estimation is obtained by testing it on part or all of samples [2]. For real applications data, only a limited instances (examples) set is available. So, if all these data are been used as training set that would cause the inability of the model to generalize to new data and the estimated error rate (accuracy) will be overly hopeful [3]. For that, here a need appears to use methods that making the best use for the limited available data for both training and testing the performance (accuracy estimation) of the used classifier. The next section will include a brief description of such methods.

The motivation of this work is to exploit the advanced level of the accuracy in the results of Artificial Intelligent algorithms; so trying to employ that proficiency in the medical domain especially in the more sensitive field "cancer diagnosis". And to increase the diagnostic accuracy of image processing

and machine learning techniques for optimum classification among normal, malignant and benign abnormalities in digital mammograms by reducing the number of misclassified cancers by comparing and selecting the best-used methods through the test and estimates those methods accuracy.

II. SUPPORT VECTOR MACHINE (SVM) CLASSIFIER

The classifier is a method takes a new data (input) as unclassified (unlabeled) instance or feature values of an observation and identifies to which class (category) it belongs. The most classifiers common uses statistical inference for categorizing a given sample with a proper label. The classifier will evaluate the presented evidence and conclude the decision in regard with the object's class that being assigned to, regardless the features' values are within or out of that class's tolerances. This methodology is using in classifying lesions as benign, malignant or normal.
Support Vector Machine could be considered as the most dominant classifiers in the machine learning. SVMs are being used widely in various applications. Like the estimation of power, in the weather prediction, the defects classification, medical diagnosis, handwriting identification, audio processing and also speaking recognition [2].

A. Linear SVM

For linearly separable problems, it finds the optimal separating hyperplane by maximizing the margin, the perpendicular distance across the hyperplane to the closest instances (the support vectors) on either side of it [5]. The examples closest to the hyperplane are called the support vectors, and the (classification) margin of the separator is the distance between support vectors from the different classes (Fig.1) [2].

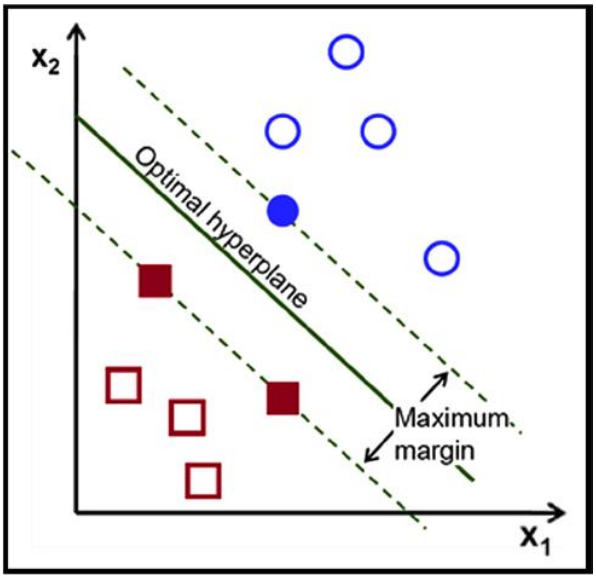

Fig.1: Classification Margin and Support Vectors in SVM.

B. Multiclass SVM

SVMs are basically two-class classifiers. The early extensions of the SVM binary classification to the multi-class case. Numerous strategies had been devised by the researchers to address the multi-classification problems [3]. One of these strategies is one-against-all (OAA) it is also called (One-versus-the-rest) which is the earliest SVM multiclass implementation and one of the most commonly used of multiclass SVMs. It constructs c binary SVM classifiers, where c is the number of classes. Each classifier distinguishes one class from all the others, which reduces the case to a two-class problem. The c decision functions can be presented as in (1):

$$w_1^T \varphi(x_i) + b_1; \dots; w_c^T \varphi(x_i) + b_c \qquad (1)$$

The formulation of the OAA method will assign the data points to the class (2). that has the maximum value, nonetheless of the sign [2]. The final label output is given to the class that has established the upper output value:

$$class\ of\ x \equiv argmax_{i=1,\dots,c}(w_i^T \varphi(x) + b_i) \qquad (2)$$

III. CLASSIFIER ACCURACY ESTIMATION

For purposes of evaluating the classification systems' performance, the performance of the used classifier should be measured. During the test phase, the binary classifier's performance is generally quantified by its accuracy i.e.,(the fraction of misclassified samples on the test set). The usage of sensitivity (true positive rate) of the classifier and its accuracy could indicate and evaluate its performance, by giving its performance (FP) false positive or (FN) false negative instances. (Table I) illustrate how the confusion matrix states the relationship between

different performances' indication for binary classification [6].

TABLE I: RELATION BETWEEN, FN, FP, TN & TP OF THE CONFUSION MATRIX

Confusion Matrix	Positive (p^a)	Negative (Na)
Positive (p^p)	True Positive (TP)	False Positive(FP)
Negative (N^a)	False Negative (FN)	True Negative(TN)

The accuracy represents the ratio of the total numbers of what is being classified correctly to the total test set [2] [6], it is given by:

$$\text{Accuracy} = \frac{\text{correctly classifiedsamples}}{\text{Total instances}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Sensitivity is also known as the True Positive Rate and is defined as the following:

$$\text{Sensitivity} = \frac{\text{Positives correctly classified}}{\text{Total positives}} = \frac{TP}{TP+FN} \quad (4)$$

IV. TESTING METHODS

There are many different methods can exploit for this task, such as [4]:

A. The Holdout Method

This is the simplest kind of validation methods. One third of the data is holdout to be used for the test by separating the whole dataset into two sets are the training set and the test set as illustrated in (fig. 2). The classifier will trains (learns) on the training data part, while its performance is estimated on the test data part. The training set usually two-third or it can be one-half from all the available data [7] [3].

There are some limitations characterized with the holdout method, the estimated accuracy which computed from the test set (smaller set) will be poor if the data part which specified for training was too large; the whole data could not use for training only. In addition to that, the test set and the training set are not independent of each other; the class which not represented in the first subset will be over represented in the other set [8].

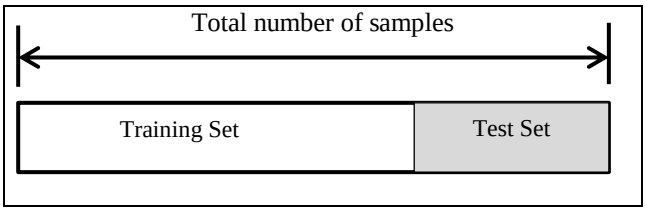

Fig. 2. The holdout method

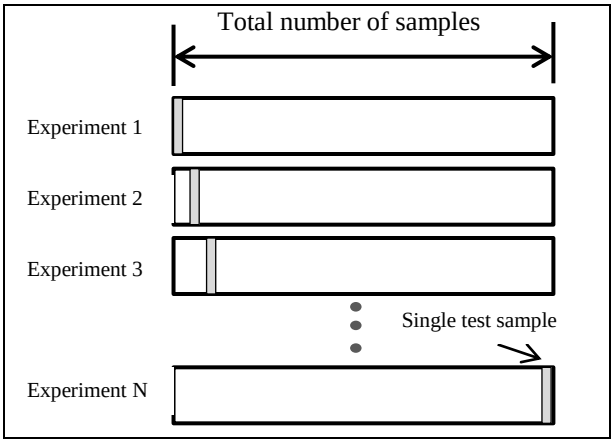

Fig. 3. Leave-one-out cross-validation

B. Cross-validation method

It is a common general method for classifier evaluation. Part of data is removing aside before training the classifier on these data. The removed part considered as a "new" data is then used for testing the performance of the classifier that has been learned.

One of cross-validation method sets is the one-leave-out approach, it considered as a special case of cross-validation, K=N, where N is the size of data set, (fig. 3) illustrates the concept of this method. In one-leave-out method each test set there is only one sample has been left where all other samples are used for training. As much data as possible uses for training in this method. The most use of this ml times then takes the average of the experiments estimates [2].

V. EXPERIMENTS

Each constructed classification system should include a test phase for evaluating the performance of the classifier that is used in the algorithm of the proposed system. In this work; 100 image samples are taken from "MIAS" dataset for the experiment. After passing these images through the used SVM classifier, two test classifier performance methods are implemented separately, a one-leave- out method is applied to estimate the used classifier error rate "accuracy" by applying the mentioned in (3) after each sample classification in the test set. at each time 99 samples are used for training and only one sample

will be used for the test, implementing that on the whole data set samples and take the average of the total (N=100) estimates. That method gives a good accuracy rate with a small misclassification ratio. But trying to increase the test samples number; another method is applied, it is the "hold-out" method. The used proportion of the training set from all the used samples is two-thirds and the rest are used for estimating the accuracy of classifying the test set. Both used estimation methods gave good results with small differences among their results in accuracy, sensitivity and time computation cost.

For the proposed computerized classification diagnoses of normal, malignant and benign, the six performance states (TN, TM, TB, FN, FM and FB) in (Table II) are stated where comparing the gained learning machine's output to the real labels that determined by a biopsy.

VI. RESULTS AND CONCLUSION

In this work, implementing multi-class SVM have been classified the segmented suspected regions "masses and/or calcifications" into normal, malignant or benign; according to statistical measurements. The SVM classifier gives good diagnoses results. After classification step has been accomplished, estimating the classifier's performance has done using two different methods, it is tested with "holdout validation" method using 70% from the samples for training and the remaining 30% for test, it obtained accuracy rates as 0.9571 % for the training set and 0.9333% for the test set; but when the 'one-leave-out validation' method was used for the same purpose and on the same selected samples, the average accuracy rate reached 0.9433 %. Another performance measurement criterion calculated which is the true positive rate "sensitivity", the one-leave-out method is overcomes the hold out method. As a comparison between these used methods in addition to their difference in the accuracy error rate, sensitivity and the running time consumed with each method are calculated and those differences ratios are listed in (table III).

TABLE II: TRIPLE CLASSIFICATION. PERFORMANCE MEASUREMENTS

Performance Measure	Definition
True normal (TN)	Tumor marked as normal by a biopsy, which is also classified as normal by the learning machine.
False normal (FN)	Tumor marked as normal by a biopsy, but is classified as benign or malignant by the learning machine.
True Benign (TB)	Tumor marked as benign by a biopsy, which is also classified as benign by the learning machine.
False Benign (FB)	Tumor marked as benign by a biopsy, but it is classified as normal or malignant by the learning machine.
True malignant (TM)	Tumor marked as malignant by a biopsy, which is also classified as malignant by the learning machine.
False malignant (FM)	Tumor marked as malignant by a biopsy, but it is classified as normal or benign by the learning machine.

TABLEIII: COMPARING ACCURACY AND TIME FOR THE TWO TEST CLASSIFIER'S PERFORMANCE METHODS

Method	Total accuracy	Sensitivity	Running time
Hold out	0.9571%	0.80%	16.2041second
One leave out	0.9333 %.	0.909%	44.4371second

From the obtained experiments results can conclude that, in comparison; in spite of gaining smaller true positive rate using hold out method than the other method; the first method recorded higher total accuracy rate in addition to costing less time comparatively.

Computing classifier accuracy using the first mentioned method with enough subset selection would be highly useful in this work, as many samples having considerably discriminative information about the abnormalities that are granted extremely robust features for the tested sample. but allocating only one sample for the test may not carry enough discriminative features to be distinguished from other class's features and classifying only one sample against other training samples may put the one test sample in intersection features area.

The proposed system's accuracy rate that tested using the hold out method is reached to 0.9571% which is higher than the previous related work in the same research field as that illustrated in (Table IV).

TABLE IV: ALGORITHM ACCURACY COMPARISON WITH RELATED WORK.

work	Total Accuracy Rate
tissue classification based on intelligence computing model [9]	93.4 %
automatic detection of breast cancer in mammogram images [10]	91.27%
texture analysis for mass classification in mammograms [11]	85.96%

mammogram image enhancement, mass segmentation and classification [12]	90.7%
Computer aided detection system for micro-calcifications in digital mammograms [13]	83%

For that, the other estimation method has been giving better estimation accuracy in less time consumed that because the one-leave-out method is computationally cost since it repeats the experiments N times (N is the number of samples).

REFERENCES

[01] Eltoukhy, Mohamed Meselhy, Ibrahima Faye, and Brahim Belhaouari Samir. "A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation." Computers in biology and medicine 42, no. 1 (2012): 123-128.

[02] Awad, Mariette, and Rahul Khanna. Efficient learning machines: theories, concepts, and applications for engineers and system designers. Apress, 2015.

[03] Dougherty, Geoff. Pattern recognition and classification: an introduction. Springer Science & Business Media, 2012.

[04] Ahuja, Yashima, and Sumit Kumar Yadav. "Multiclass classification and support vector machine." Global Journal of Computer Science and Technology Interdisciplinary 12, no. 11 (2012): 14-20.

[05] Alexandre Kowalczyk. Support vector machines succinctly released. Syncfusion, 2017.

[06] Felkin, Mary. "Comparing classification results between n-ary and binary problems." In Quality Measures in Data Mining, pp. 277-301. Springer, Berlin, Heidelberg, 2007

[07] Dobbin, Kevin K., and Richard M. Simon. "Optimally splitting cases for training and testing high dimensional classifiers." BMC medical genomics 4, no. 1 (2011): 31.

[08] Baumann, Désirée, and Knut Baumann. "Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation." Journal of cheminformatics 6, no. 1 (2014): 47.

[09] Mustafa nafea, "tissue classification based on intelligence computing model", Mustansiriyah university, 2017.

[10] Faozia A. S. alsarori "automatic detection of breast cancer in mammogram images", 2013

[11] Y. LI ET AL. "texture analysis for mass classification in mammograms", pattern recognition letters 52,2014, 87-93.

[12] N. AL-NAJDAWI ET Al, "mammogram image enhancement, mass segmentation and classification", applied soft computing 35, 2015, 175-185.

[13] H. Mohamed, M.S. mabrouk, A. sharawy, "Computer aided detection system for micro-calcifications in digital mammograms", computer methods and programs in biomedicine, 2014.