2011 Eighth International Conference Computer Graphics, Imaging and Visualization

A Comparison of SVM Kernel Functions for Breast Cancer Detection

Muhammad Hussain[1], Summrina Kanwal Wajid[2], Ali Elzaart[1], Mohammed Berbar[1]
Department of Computer Science, King Saud University, Riyadh, KSA[1]
Alyamamah University, Riyadh, KSA[2],
{mhussain@ksu.edu.sa, summrina@gmail.com}

Abstract

Support vector machines outperform other classification methods for breast cancer detection. However the performance of SVM is greatly affected by the choice of a kernel function among other factors. This article presents a comparative study of different kernel functions for breast cancer detection. The focus is on classification using SVM with different kernel functions. The comparison with neural network based method using MLP is also given. Furthermore, we examine the affect of selecting feature subsets before applying classification with different kernels. For features subset selection we used genetic algorithm. The evaluation is based on 5 X 2 cross validation.

1. Introduction

Breast Cancer is the second major cause of deaths in women all around the world. A woman living in the United States has a 12.28% or 1 in 8 probability of developing invasive breast cancer sometime during her lifetime. In Asia the death rate due to breast cancer is increasing dramatically. The problem in Asia pacific is that women do not go for check up, and so early diagnosis is not possible. There is a great need to come up with early diagnostic system for breast cancer detection since it can minimize the risk of death. CAD can help in early detection. It can help doctors in making accurate recommendations for a current case.

Many techniques are being used to predict and classify breast cancer pattern. Mammography is one of the commonly used methods to detect the breast cancer but radiologists show variation in interpreting mammographic image. There is a need to have an automated method to interpret the mammographic image. An automatic method is based on some classifier. Different classifiers have been employed for this purpose and there has almost been a consensus on that SVM (support vector machine) outperforms other classification methods for breast cancer detection. However the performance of an SVM classifier is greatly dependent on the proper choice of a kernel function among other factors. This paper gives a comparison between different kernel functions for breast cancer detection. We used four typical kernels and showed the performance of each of them. Further, we investigated the affect of these kernels with feature subsets selected using genetic algorithm. This scheme helps in extracting the best feature subset that can improve classification accuracy and save computation time.

The organization of the rest of the paper is as follows. Section 2 gives an overview of the related work. Section 3 presents the methodology. Section 4 describes the evaluation framework. The measures used for evaluation have been discussed in Section 5. Section 6 presents the comparative results. Discussion and conclusion have been given in Sections 7 and 8.

2. Related Work

There is a continuous study and research going on in this field. One of the methods for interpreting the mammographic image suggested in [1] uses Multi-Layer Perception (MLP) neural network based classification. It first reduces the feature vector dimensions which were obtained using association rules. Next Multi-Layer Perception (MLP) is used for the classification of the features obtained. In [12], the author combined different techniques like multilayer perception, support vector machine and combined neural network, probabilistic neural network, recurrent neural network for classifying the data.

Use of different kernel functions in SVM results in different performance. In [2] the authors have used two different kernel functions, polynomial kernel function and Gaussian radial basis function. Performance of classification was analyzed using overall classification accuracy, confusion matrix, sensitivity and specificity measures. For polynomial kernel functions, the 4th degree polynomial function gives overall best performance of 92.627% (sensitivity = 92.69%, specificity = 92.564%). In case of radial basis functions, the highest overall accuracy (92.105%) was obtained with Sensitivity = 93.252% and specificity = 91.018%.

In [3], the authors propose two-class SVM classification for detection of masses in mammogram images. This technique uses Moran's index and Geary's coefficient as feature vectors to be classified using radial basis as a kernel function. A content-based mammogram retrieval system for diagnosis of micro-calcifications

978-0-7695-4484-7/11 $26.00 © 2011 IEEE
DOI 10.1109/CGIV.2011.31

(MCs) was proposed in [4]. The scheme presented here is a modification of SVM. The classification accuracy of different classifiers, including MLP and SVM on Wisconsin dataset has been compared in [12].

In [14], SVM has been used for detection of micro-calcification clusters in digital mammograms. It uses polynomial and radial basis function as kernels in SVM. SVM with radial basis function has also been used for classification of the malignant masses in [14]. But before applying classifier they found the best values of RBF model (C, γ) by applying cross validation. In [15] author has used Least Square Support Vector machine which uses linear equation for training unlike simple SVM, with k-fold cross validation. It uses RBF to transform non-linear data to higher dimensional feature space. Using WDBC dataset with 9 features it got accuracy of 98%. SVM has also been used in [6], and [7] for classification purpose.

3. Methodology

We evaluate the state of the art classification techniques SVM and neural network for breast cancer detection problem.. For SVM we use different kernels, and compare their relative accuracy. In case of neural network, we use Multilayer Perception (MLP) network. We also compared the affect of features subset selection on the accuracy of classification schemes. Each method has been evaluated using 5 X 2 cross validation scheme. The feature subset selection technique is inspired by the method suggested in [8] using genetic algorithm.

3.1 Feature subset Selection

Genetic algorithm is based on the theory of evolution. It is used for solving the optimization problems. It starting with the initial random population of solutions, where each solution is represented by a chromosome. Based on certain criteria new generation of solutions is formed from previous generation and are assumed to be better than parent solution. This process is repeated several times until a certain condition is met.

3.1.1 Initial Population

Each solution is represented as a fixed length chromosome bits, where each bit corresponds to a feature in a feature vector. The basic algorithm is as followed:

1. Create initial population.
2. Evaluate the population applying certain function
3. Select new population from the old one applying mutation and cross-over
4. Evaluate new population
5. Repeat 2 - 4 until certain criteria is met [9] We experimented with different sizes of initial population ranges from 100-300. In a chromosome, the presence of the bit 1 means that the corresponding feature will be selected.

3.1.2 Evaluation

Initial population is evaluated to find the best population which can be mutated. For that purpose the criteria is to have minimum number of features that can result in highest accuracy.

3.1.3 Selection

New best fit population is selected to survive in next generation. We use roulette wheel selection. In this selection method, probability of each individual is calculated. The individual with higher probability has more chance to be selected.

3.1.4 Cross Over

We have tested different cross-over operators: uniform, one-point, and arithmetic. The detail of these operators can be found in [9].

3.1.5 Mutation

Cross-over operator confines the search is a local region and the may be trapped in a local minima. Mutation operation is a solution to this problem. A simple mutation operator which flips random bits with probability of 0.04 is used.

3.2 Classification

We used Support Vector Machine with different kernel methods and MLP for breast cancer detection. In the following sections, we give an overview of these methods.

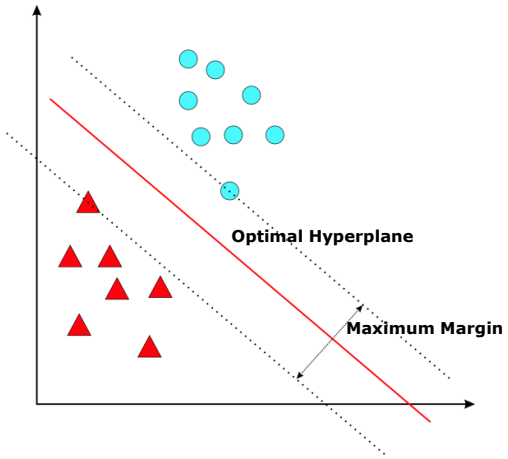

Figure 1: SVM classifies by finding the optimal hyperplane that has maximum margin.

3.2.1 SVM

Support Vector Machine (SVM) is part of a group of kernel based methods which are used for pattern classification and regression. A classifier takes an input pattern called feature vector, and determines to which class it belongs to.

Let x_i, i =1, 2, …, M be feature vectors of a training set X, which belong to either of two classes ω_1 and ω_2. Using this training data, SVM finds an optimal hyper plane

$$g(x) = w^T x + w_0 = 0 \quad (11)$$

with maximum margin that separates the unknown input patterns into 2 classes as shown in Figure 1. Many hyperplanes separating the feature vectors are possible, SVM finds the one that has maximum margin and better generalization performance for classification.

SVM is basically a linear classifier that classify linearly separable data, but in general , the feature vectors might not be linearly separable. To overcome this issue, kernel trick is used. The original input space is mapped into a high-dimensional feature space using kernel functions where it becomes linearly separable. The performance of an SVM classifier is dependent on the choice of a proper kernel function. Different kernel functions have been employed for different classification tasks. We employ four kernels functions (polynomial, radial basis function, Mahalanobis, and sigmoid) for breast cancer detection and compare their performance.

- Polynomial kernel with degree d can be written as
$$\mathrm{K}(x_i, x_j) = \left(\gamma {x_i}^{\mathrm{T}} x_j + 1\right)^d, \gamma > 0$$
- Radial basis function kernel is given as:
$$\mathrm{K}(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \gamma > 0$$
- Sigmoid kernel function is given as
$$\mathrm{K}(x_i, x_j) = \tanh\left(\gamma {x_i}^{\mathrm{T}} x_j + r\right)$$
It is also called known as Multi Layer Perception Kernel or Hyperbolic Tangent Kernel.
Here
γ, r and d are adjustable kernel functions which are adjusted based on the data.
- The Mahalanobis kernel is approximated as
$$\mathrm{K}(x_i, x_j) = -\frac{\delta}{m}\left(x_i - x_j\right)^{\mathrm{T}} \mathrm{Q}^{-1}\left(x_i - x_j\right)$$
where δ (> 0) is the scaling factor to control the Mahalanobis distance and Q is the covariance matrix for given dataset and *m* is the dimension of the data set.

3.2.2 MLP

Neural Network is made up of interconnected artificial neurons. They mimic human brain processing. The neurons interconnection link carries certain weight. The output of each neuron is determined by using an activation function such as sigmoid and step. In case neural networks (NN) are trained with training pattern of known classes, these are called supervised learning NN.

The supervised learning process of the neural network consists of a unique input signal and corresponding desired output signal .The network is trained until it reaches a stable state where the synaptic weights doesn't change and maps to their corresponding output. In recent years, there had been a great research in using neural network for classification of the mammography images.

In case of multi layer perception (MLP), neurons are connected in a network topology. They are placed in different layers and are connected through certain weights. We use 3 layer MLP containing input, hidden and output layers. The input layer consists of as many neurons as the number of features in a feature vector. Second layer, called *hidden layer,* contains *h* number of perceptions, where value of *h* is determined by experiment. Output layer contains only one neuron representing either benign or malignant value. We used sigmoid activation function for hidden and output layers. Batch learning method is used for updating weights between different layers.

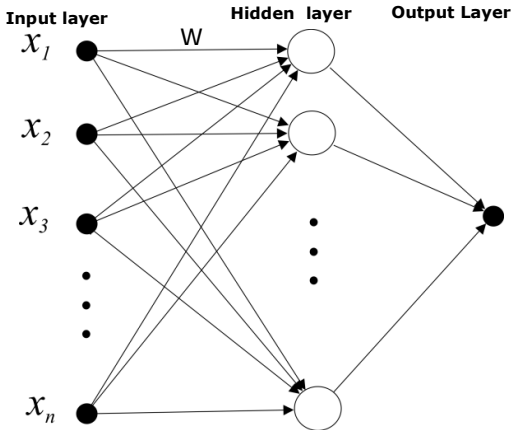

Figure 2 Architecture of a Multilayer Perception

4. Database for Evaluation of the Methods

We use Wisconsin breast cancer database from Machine Learning Repository of the University of California, for the comparative study. It contains 699 feature vectors out of which 458 are benign cases and 241 belong to malignant class; 16 feature vectors are incomplete. We discord the incomplete records from our experiments. Each feature vector contains 9 features. The value of each feature ranges from 1 - 10, where 10 means the most abnormal state. The features are listed below:

(i) Thickness of clump.
(ii) Cell size uniformity.
(iii) Cell shape uniformity.
(iv) Marginal adhesion.
(v) Single epithelial cell size.
(vi) Bare nuclei.
(vii) Bland chromatin.
(viii) Normal nucleoli.
(ix) Mitoses.

Performance of classification with SVM employing different kernel functions has been evaluated using 5 x 2 cross validation method. It performs 5 replications of 2-foldcross validation. In each replication, dataset is randomly portioned into 2 equal sized sets. Further learning algorithm is trained on one of the sets, and then tested on the other set. Performance is given in terms of sensitivity, specificity, and overall accuracy.

5. Measures of Performance

We employ commonly used measures of performance: overall accuracy, sensitivity and specificity.

Sensitivity is a measure of accuracy of diagnosis of true (malignant) cases. It is defined as:

Table 1 Classification results with SVM using different kernel functions.

Kernel Type	*#features*	*logC*	*logγ*	*d*	*logr*	*logδ*	*Sensitivity*	*Specificity*	*Accuracy*
RBF	*9*	3.5	-12.5				95.14247	96.88576	96.28063
	5(ii,iii, iv, vi,ix)	-1.5	-11				94.7187	97.52745	96.54542
	4(ii, v, vi, viii)	5	-15				96.22	96.391	96.33886
	4(ii, iv, vi, viii)	1.5	2.5				94.644	97.022	96.193
	3(ii, vi, viii)	6.5	-15				96.52376	96.89481	96.74967
Poly	*9*	1.5	-0.5	4			99.14919	96.63809	97.51119
	5(ii,iii, iv, vi,ix)	3	-3	4			96.90117	96.5401	96.66229
	4(ii, v, vi, viii)	-0.5	-5	4			98.73245	95.95026	96.92519
	4(ii, iv, vi, viii)	3.5	-4	4			97.02024	96.43672	96.632
	3(ii, vi, viii)	5	-4	4			97.074	96.57	96.75001
Sigmoid	*9*	2.5	-12		-0.6		95.39694	97.87456	97.01351
	5(ii,iii, iv, vi,ix)	2	-10		0		92.62661	97.87727	96.0468
	4(ii, v, vi, viii)	1.5	-11.5		0		93.56842	97.54626	96.16445
	4(ii, iv, vi, viii)	-3.5	-14.5		-1.8		91.65	98.372	96.016
	3(ii, vi, viii)	1.5	-6		2.4		93.33225	97.66704	96.13503
Mahalanobis	*9*	1.1				1.1	96.63	97.29	97.06
	5(ii,iii, iv, vi,ix)	1.1				1.1	94.12	97.75	96.48
	4(ii, v, vi, viii)	0.1				0.1	95.79	95.94	95.89
	4(ii, iv, vi, viii)	1.1				1.1	95.79	95.94	95.89
	3(ii, vi, viii)	1.1				1.1	92.43	97.74	95.89

$$Sensitivity = \frac{TP}{TP + FN}\%$$

Specificity is a measure of accuracy of diagnosis of false (benign) cases. It is defined as:

$$Specificity = \frac{TN}{FP + TN}\%$$

where

TP: *True Positive*: A patient predicted with breast cancer when he/she actually has breast cancer,

TN: *True Negative*: A patient predicted healthy when he/she actually is healthy,

FP: *False Positive*: A patient predicted with breast cancer when he/she actually is healthy, and

FN: *False Negative*: A patient predicted healthy when he/she actually has breast cancer.

Overall *accuracy* is given as

$$Accuracy(T) = \frac{\sum_{i=1}^{T} Assess(t_i)}{|T|},\ t_i \in T$$

where *T* is the set of test samples and

$$Access(t) = \begin{cases} 1\ if\ t\ is\ correctly\ classification \\ 0\ if\ t\ is\ incorrectly\ classification. \end{cases}$$

6. Experiment Results

We use all 9 features as well as subsets of the features selected using GA algorithm for evaluating the performance of different kernel functions. Table 1 lists the classification results obtained using SVM classifier with different kernel functions. We use different parameters for GA algorithm, which result in different subsets of features. Note that we select the subsets of features using only RBF kernel and GA with different sets of parameters and use these subsets to measure the classification performance of other kernels. The reason is to use the same feature subsets for fair comparison. Each kernel function has parameters and the performance of SVM depends on proper tuning of these parameters. We use coarse and fine grid search and 5x2 cross-validation to find the optimal values of these parameters. These values are given in Table 1. Please note that in this table $\log_2$ of these parameters is given.

We select a subset of features using GA with population of size 100, 200 generations and fitness convergence as stopping criterion. For selection, we employ four methods: top 20% , best, tournament, and random; all these methods lead to the same result. Uniform crossover operator with 0.66 cross-over probability, and mutation with 0.04 as mutation probability were used. This instance of GA selects the following subset consisting of 5 features:

(ii) Cell size uniformity.
(iii) Cell shape uniformity.
(iv) Marginal adhesion.
(vi) Bare nuclei.
(ix) Mitoses

In the second instance of GA, we keep the population size equal to 100, the number of generations equal to 300 and the stopping criterion is fitness convergence. For selection, we use roulette wheel method. We employ uniform cross-over operation with 0.9 cross-over probability and mutation operation with mutation probability being 0.04. In this case, the following subset of features is selected:

(ii) Cell size uniformity.
(v) Single epithelial cell size.
(vi) Bare nuclei.
(viii) Normal nucleoli.

Next we select another feature subset employing GA with population of size 100 and stopping criterion being 300 generations with fitness convergence. Selection method was choosen to be roulette wheel with uniform cross-over operator with cross-over probability of 0.66. Mutation operator with 0.04 mutation probability was used. The features selected in this case are given below:

(iii) Cell size uniformity.
(iv) Marginal adhesion.
(vi) Bare nuclei.
(viii) Normal nucleoli.

Finally, we select a subset of features with GA that employs a population of size 100 and stopping criterion being 300 generations with fitness convergence. For selection, roulette wheel method was used. For cross-over and mutation operators, two-point operator with probability of 0.66 and mutation rate of 0.04 were used. In this a feature subset of three features is selected, which are given below:

(ii) Cell size uniformity.
(vi) Bare nuclei.
(viii) Normal nucleoli.

In case of MLP, for measuring the recognition performance, we use all 9 features and feature subset selected with GA. For MLP, we applied GA with population of size 50 and 100 generations for feature subset selection and came up with following features:

(iii) Cell shape uniformity.
(iv) Marginal adhesion.
(v) Single epithelial cell size.
(vi) Bare nuclei.
(viii) Normal nucleoli.

The recognition results with all 9 features and the selected feature subset are shown in Table 2.

Table 2 Classification results with MLP

Performance	With all features	With selected features
Accuracy	94.63%	94.66%
Sensitivity	94.44%	93.56%
Specificity	95.614%	95.01%

7. Discussion

A close look at Table 1 reveals that in case of RBF kernel the classification rate improves when a feature subset is used in place of full set of features. In case of other kernels, there is a small deterioration in recognition. This can be due to the reason that we selected the feature subsets with RBF. In general, it can be concluded that selecting best feature subset can enhance the classification accuracy. Also, it is interesting to note that the results of SVM employing polynomial kernel of degree 4 and minimum number of features are better than other kernel function even with all the features.

In case of Mahalanobis kernel function subset selection does not affect the performance but we can say that we came up with the set of features which contribute towards better classification as compare to using all of them. In case of sigmoid kernel function again selecting a proper subset of features can enhance the classification rate.

The performance of MLP for classification is less than that with SVM using different kernel functions. Also, it is clear from Table 2, feature subset selection does not significantly affect the performance.

8. Conclusions

We have presented the results of comparative study on SVM with different kernels for breast cancer detection. For comparison, we employed four kernels: RBF, polynomial, Mahalanobis, and sigmoid. We evaluated the affect of these kernels with and without feature subset selection. This study indicates that sigmoid results in the best specificity whereas polynomial function gives the best sensitivity even after feature subset selection. Further this study indicates that SVM with kernel having worse performance is better than MLP. It follows from the results that an SVM classifier based on a combination of kernels can further enhance its performance for breast cancer detection. We will further investigate other kernels. Also, investigation on how to combine different kernels for better detection results is our future work.

Acknowledgment

This work is supported by the National Plan for Science and Technology (NPST), King Saud University, Riyadh, Saudi Arabia under the project 08-INF325-02.

References

[1] Murat Karabatak, Elazig, M. Cevdet Ince. "An expert system for detection of breast cancer based on association rules and neural network", *Journal Expert Systems with Applications: An International Journal archive*, Volume 36 Issue 2, March, 2009

[2] Muthu Rama Krishnan, Shuvo Banerjee, Chinmay Chakraborty, Chandan Chakraborty, Ajoy K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine" , *Expert Systems with Applications 37*, page: 470-478, 2010.

[3] Sung-Nien Yua, Kuan-Yuei Lib, Yu-Kun Huangac, "Detection of micro calcifications in digital mammograms using wavelet filter and Markov random field model", *Computerized Medical Imaging and Graphics* ,Volume 30, Issue 3, Pages 163-173, April 2006.

[4] Yongyi Yang, Liyang Wei, Nishikawa. R.M., "Micro calcification Classification Assisted by Content-Based Image Retrieval for Breast Cancer Diagnosis", *ICIP 2007, IEEE International Conference*, Volume: 5, page(s): V - 1 - V - 4.

[5] Mehmet Fatih Akay, "Support vector machines combined with feature selection for breast cancer diagnosis", *Expert Systems with Applications*, 2009.

[6] Alfonso Rojas Domínguez, Asoke K. Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms", *Pattern Recognition 42*, 2009.

[7] Waei A. Mohamed, Mohamed A. Alolfe, Yasser M. Kadah. "Fast Fractal Modeling of Mammograms for Micro calcifications Detection", *26th NATIONAL RADIO SCIENCE CONFERENCE (NRSC2009)*, 2009.

[8] Zehang Sun, George Bebisa, Ronald Miller, "Object detection using feature subset selection ", Pattern Recognition 37 (2004) , pp. 2165 – 2176.

[9] S. Sumathi, Surekha Paneerselvam. "Computational Intelligence Paradigms Theory & Applications using MATLAB", CRC Press; 1 edition, January 5, 2010.

[10] Sergios Theodoridis, Dr. Aggelos Pikrakis., Konstantinos Koutroumbas, Dionisis Cavouras., "Introduction to Pattern Recognition: A Matlab Approach", Academic Press, 2010.

[11] Shigeo Abe. "Support Vector Machines for Pattern Classification", Springler.

[12] Elif Derya Ubeyli, "Implementing automated diagnosis systems for breast cancer detection", Expert Systems with Applications, pp:1054-1062,2007.

[13] Issam El-Naqa, Yongyi Yang, , Miles N. Wernick, "A Support Vector Machine Approach for Detection of Microcalcifications", *IEEE TRANSACTIONS ON MEDICAL IMAGING*, Vol. 21, No. 12, December 2002.

[14] Cheng Lung Huang, Hung Chang Liao, Mu Chen Chen, " Prediction model building and feature selection with support vector machines in breast cancer diagnosis", Expert Systems with Applications 34 (2008) 578–587.

[15] Kemal Polat, Salih Güne¸s, "Breast cancer diagnosis using least square support vector machine", Digital Signal Processing 17 (2007) 694–701.