www.tj-es.com

Tikrit Journal of Engineering Sciences/Vol.20/No.5/June 2013, (43-49)

Breast Tumor Classification Using SVM

Joanne H. Al-Khalidy
Assist. Lecturer
Electronics Engineering College
University of Mosul

Raid R. Al-Ne'ma
Lecturer
Technical Computer Engineering
Technical college of Mosul

Received 19 February 2012; accepted 12 November 2012
Available online 18 July 2013

Abstract

Although there are several techniques that have been used to differentiate between benign and malignant breast tumor lately, support vector machines (SVMs) have been distinguished as one of the common method of classification for many fields such as medical diagnostic, that it offers many advantages with respect to previously proposed methods such as ANNs. One of them is that SVM provide a higher accuracy, another advantage that SVM reduces the computational cost, and it is already showed good result in this work.

In this paper, a Support Vector Machine for differentiation Breast tumor was presented to recognize malignant or benign in mammograms. This work used 569 cases and they were classified into two groups: malignant (+1) or benign (-1), then randomly selected some of these samples for training model while others were used for test. The ratios were 0.4348% of accepted false, 2.1739% of refused false. These results indicate how much this method is successful.

Keywords: SVM, Breast tumor, Classification, Benign, Malignant.

تصنيف أورام الثدي باستخدام SVM

الخلاصة

على الرغم من التقنيات العديدة التي ظهرت مؤخراً في التمييز بين المرض الخبيث عن الحميد، فإن تقنية آله المتجه المدعم (SVM) تم إعتمادها كواحدة من طرق التمييز في عدة مجالات كالتشخيص الطبي، حيث تقدم عدة فوائد منها، إن آلة متجه الدعم تعطي دقة عالية، أيضا فإنها تقلل من كلفة الحسابات، ولقد أعطت نتائج جيدة في هذا البحث.

في هذا البحث، تم تمييز الورم الخبيث عن الحميد في أورام الثدي من خلال إستخدام تقنية آلة متجه الدعم حيث تم استخدام 569 حالة نصفها للتدريب ونصفها الآخر للفحص، إذ تم إعطاء قيمة لكل حالة سليمة مساوية (+1) وكل حالة غير سليمة مساوية (−1) فكانت نسبة خطأ القبول 0.4348% ، أما نسبة خطأ الرفض فكانت 2.1739% مما يدّل على دقّة النتائج المستحصلة.

Introduction

The normal body cells grow, divide, and die for the entire life of a person. At the age of maturity, the growth of new cells is limited to replacement of dying cells or damage cells to repair [1]. Body cells sometimes continue to grow and divide abnormally. These cells may join together to form a mass of extra tissue called a tumor. There are two types of tumors: benign and malignant. Benign tumor is a non-malignant, non-cancerous tumor. It is usually localized, rarely spreads to other parts of the body and responds well to treatment. However, if left

untreated, benign tumors may lead to serious disease.

Malignant tumor is cancerous. It is resistant to treatment, may spread to other parts of the body and often recurs after removal. One of the most wide spread types of tumor is breast cancer.

Breast cancer is a "malignant neoplasm of the breast." A cancer cell has characteristics that differentiate it from normal tissue cells with respect to: the cell outline, shape, structure of nucleus and most importantly, its ability to spread and infiltrate. When this happens in the breast, it is commonly termed as 'Breast Cancer'. Cancer is confirmed after a biopsy (surgically extracting a tissue sample) and pathological evaluation [2].

The recognition of breast tumor is one of the most important factors in recovery from disease and to find a type of the tumor which can be malignant or benign.

Many researches have been done in this field. A. Marcano-Cedeno, J. Quintanilla-Dominguez, and D. Andina introduced approach in neural network training for pattern classification in breast cancer in [3]. Yulei Jiang, Robert M. Nishikawa, Dulcy E. Wolverton, Charles E. Metz, Robert A. Schmidt, Kunio Doi presented an automated computerized classification scheme to classify malignant and benign clustered microcalcifications in mammograms in[4]. Hyunsoo Kim, Peg Howland, Haesun Park introduced a decision functions for the centroid-based classification algorithm and support vector classifiers to handle the classification problem where a document may belong to multiple classes in[5]. Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, and Mark Helvie used hybrid unsupervised /supervised structure classifier and applied to classification of malignant and benign masses on mammograms in[6]. P. Babaghorbani, S. Parvaneh, AR. Ghassemi, K. Manshai choose superior individual textural features calculated from GLCM in classification of malignant or benign breast tumor in [7].

This work aim to use the Support Vector Machines (SVMs) classification algorithm to separate malignant tumors from benign ones in breast tumor.

This paper is organized as follows. In section 2, the support vector machine concept is introduced. The classification by SVM method is illustrated in section 3. Then procedure and results are shown in section 4. Finally some conclusions are presented.

Support Vector Machine Concept

The support vector machine (SVM) is an algorithm based on training and using for learning classification, introduced in 90's by Vapnik and other researchers [8]. It is based on two key concepts: A maximal margin classifier and kernel function.

A *maximal margin classifier* means that we have to look for a linear normed space. It constructs a specific hyperplane and finds an algorithm that separates all data correctly and maximizes the minimal distance between the data and the hyperplane [9]. Accordingly, the decision function for classifying points with respect to the hyperplane has to use data in terms of dot products. Thus, a support vector machine can locate a separating hyperplane in the feature space and classify points in that space without ever representing the space explicitly, simply by defining a function, called a *kernel function*, which return the inner product between two points in a suitable feature space, thus defining a notion of similarity, with little computational cost even in very high-dimensional spaces [10].

The selection of an appropriate kernel function is important, since the kernel function defines the transformed feature space in which the training set examples will be classified [11]. As long as the kernel function is legitimate, an SVM will operate

correctly even if the designer does not know exactly what features of the training data are being used in the kernel-induced feature space.

In support vector classification, the separating function can be formulated as a linear combination of kernels associated with the support vectors as:-

$$f(x) = \sum_{x_i \in S} \alpha_j y_j K(x_j, x) + b \quad \ldots\ldots (1)$$

Where x_j, $j = 1,\ldots..,m$ are the selected training examples called support vectors, and $\boldsymbol{x}$ is the input vector, $\boldsymbol{K}(\boldsymbol{x_j}\ ;\ \boldsymbol{x})$ called kernel is a symmetric positive function, $\boldsymbol{y_j}$ the label for the vector (1,-1), and $\boldsymbol{\alpha_j}$ a weight for the support vector determined in the training process, $\boldsymbol{b}$ is the bias of the hyperplane.

There are several different types of kernels as shown below:-

Linear: K (xi, x) = xi. x ……...….(2)

Polynomial: K (xi, x) = (xi. x + 1)^d …….…... (3)

Gaussian: K (xi, x) = $e^{-||\mathbf{x_i}-\mathbf{x}||^2/2\sigma^2}$ …....(4)

Where $\boldsymbol{d}$ is the degree of the polynomial kernel and $\boldsymbol{\sigma}$ the variance of the Gaussian. The polynomial and Gaussian are nonlinear kernels, which are important if there is no linear relation between the labels and the input. These kernels can solve non-linear problems [12]. Figures (1), (2) and (3) shows these types of kernels.

The Classification by SVM method

The task of classification usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes (i.e. the features or observed variables) [13]. This task produces a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

In this work, the classification was implemented upon MatLab programming language using SVM functions. The data set of breast cancer used in this paper is collected by Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison. This data set contained 569 cases with a mixture of benign and malignant masses which all are in the same file. They have separated the groups with a line beginning with ID number, Diagnosis (M = malignant, B = benign), and 30 real-valued input features. These features are computed from a digitized image of a fine needle aspirates (FNAs) of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are calculated for each cell nucleus:

1- Radius (mean of distances from center to points on the perimeter).
2- Texture (standard deviation of gray-scale values).
3- Perimeter.
4- Area.
5- Smoothness (local variation in radius lengths).
6- Compactness (perimeter^2 / area - 1.0).
7- Concavity (severity of concave portions of the contour).
8- Concave points (number of concave portions of the contour).
9- Symmetry.
10- Fractal dimension ("coastline approximation" - 1).

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were calculated for each image, resulting in 30 features [14].

10 feature mean + their standard error + their maximum =30 features.

SVM requires that each data instance is represented as a vector of real numbers. All feature values are recoded with four significant digits.

In this work the data have been addressed the classes to ±1 so that it can be analyzed by an SVM directly. We use some of this set of data for training examples, each marked as belonging to one of two groups. This training algorithm constructs a model that assigns new examples into one group or the other. This model is a representation of the examples as points in the feature space, mapped so that the examples of the separate

categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [15].

Each row of matrix of training data corresponds to cases, and each column corresponds to a feature of case. These support machines will be trained to recognize the data and separate them into malignant or benign tumor.

The procedure and results

In this work, the breast tumor classification was employed depending on the support vector machine.

The SVM algorithm adopted in this work is as follow:

SVM algorithm

1- Initialization: load the data sets, which include 32 features for each patient.

```
x=xlsread('book1.xls',2);
```

2- Create *p*, a two-column matrix containing a label for 230 cases as 'B' for benign and 'M' for malignant.

```
p=x(1:230,1:2);
```

3- From the species vector, create a new column vector as a group *t*, to classify data into two groups: -1 and +1.

```
t=x(1:230,31);
```

4- Randomly select some of cases for training sets and others for test sets.

5- Train an SVM classifier using a linear kernel function and plot the grouped data.

```
net=svmtrain(p,t,'Showplot','true','Kernel_Function', 'linear');
```

6- Stop when get minimum FAR & FRR for training set. Otherwise repeat 5.

7- Test the performance of trained SVM using set tests.

Figure (4) shows the flowchart of the proposed program, this program has been designed using MatLab which works under the system of windows.

In MatLab program, the database was loaded as an Excel file which includes 569 cases of breast cancer, each case has 32 cells of features. In this file, a new column vector (±1) has been created to classify data into two groups: malignant or benign. Then, we randomly selected some of these samples for training model while others were used for test. The linear kernel function was used in the training of SVM classifier. The plotting of the grouped data is shown in figure (5). Then some cases tested as shown in figure (6).

The following equations were used to examine our results.

far= (fa/230)*100% ……..……(5)

Where (far) means accepted false ratio, (fa) is number of accepted false, 230 is number of test cases.

frr= (fr/230)*100% ……(6)

Where (frr) means refused false ratio, (fa) is number of refused false, 230 is number of test case.

The ratios were: 0.4348% of accepted false, 2.1739% of refused false.

These results indicate how much this method is successful.

In the figure (7) shows the mean absolute error for train = 0.0261.

And in the figure (8) shows the mean absolute error for test = 0.0522.

The X-axis in figure (7) and figure (8) mean the number of cases and Y-axis means the value.

Conclusions

From results, it can be shown that the ratios were 0.4348% of accepted false, 2.1739% of refused false. These results indicate that SVM is an appropriate alternative to traditional algorithms but it has rapid convergence and unique solution. SVM algorithms have a simple geometric interpretation and give a sparse solution. As well as SVMs often outperform ANNs in practice because they deal with the biggest problem with ANNs and SVMs are less prone to overfitting. Unlike conventional statistical and neural network methods, the SVM approach does not attempt to control

model complexity by keeping the number of features small.

References

1- American Cancer Society, "Leukemia-Chronic Lymphocytic", www.cancer.org, 2011.

2- A.V. Deshpande, S.P. Narote, V.R. Udupi and H.P. Inamdar, "A Region Growing Segmentation for Detection of Micro calcification in Digitized Mammograms", Proceedings of the International Conference on Cognition and Recognition 774.

3- A. Marcano-Cedeño, J. Quintanilla-Domĩnguez, D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network" ,

4- Expert Systems with Applications 38 (2011) 9573–9579.

5- Yulei Jiang, Robert M. Nishikawa, Dulcy E. Wolverton, Charles E. Metz, Robert A. Schmidt, Kunio Doi., "Computerized Classification of Malignant and Benign Clustered Microcalcifications Mammograms", 19th International Conference - IEEE/EMBS Oct. 30 - Nov. 2, 1997 Chicago, IL. USA.

6- Hyunsoo Kim, Peg Howland, Haesun Park, "Dimension Reduction in Text Classification with Support Vector Machines", Journal of Machine Learning Research 6 (2005).

7- Lubomir Hadjiiski, Member, IEEE, Berkman Sahiner, Member, IEEE, Heang-Ping Chan, Nicholas Petrick, Member, IEEE, and Mark Helvie, "Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach", IEEE Transactions on medical imaging, Vol. 18, No. 12, December 1999.

8- P. Babaghorbani, AR. Ghassemi, S. Parvaneh, K. Manshai, "Sonography Images for Breast Cancer Texture classification in Diagnosis of Malignant or Benign Tumors", 978-1-4244-4713-8/10/$25.00 ©2010 IEEE.

9- Yiqiang Zhana,b,c, Dinggang Shenb,c, "Design efficient support vector machine for fast classification", The Journal of the Pattern Recognition Society, 38 (2005) 157–16, www.elselviercomputerscince.com.

10- Vojtêch France, Vaclav Hálváć, "An iterative algorithm learning the maximal margin classifier", the Journal of the Pattern Recognition Society, 36 (2003) 1985-1996.

11- Alexandros Karatzoglou, David Meyer, Kurt Hornik, "Support Vector Machines in R", Journal of Statistical Software, April 2006, Volume 15, Issue 9.

12- G. Garg, Vijander Singh, Mudita Grover, Nidhi, J.R.P Gupta, "Optimal Kernel Learning for EEG based Sleep Scoring System", International Journal of Biological & Medical Research, Int J Biol Med Res. 2011; 2(4): 1220 – 1225.

13- Pascal Paysan, "Stereovision based vehicle classification using support vector machines", Fachbereich Information's technique, Software technique in partial fulfillment of the requirements for the degree of Diplom-Ingenieur – Software technique, February 2004.

14- Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, April 15, 2010.

15- W.N. Street, W.H. Wolberg and O.L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

16- Support Vector Machine, http://en.wikipedia.org/wiki/Support_vector_machine.

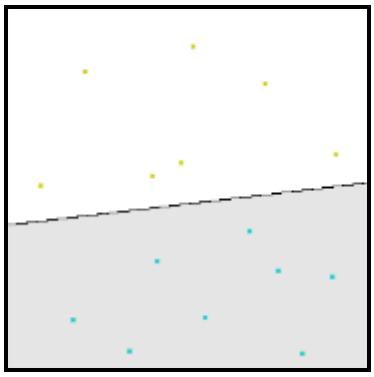

Figure (1): Linear Kernel

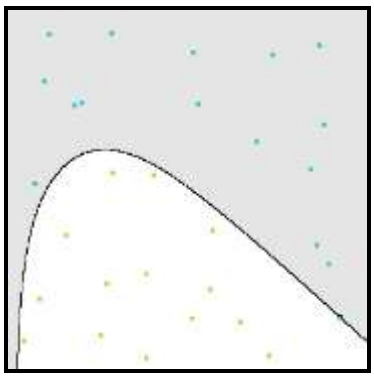

Figure (2): Polynomial Kernel

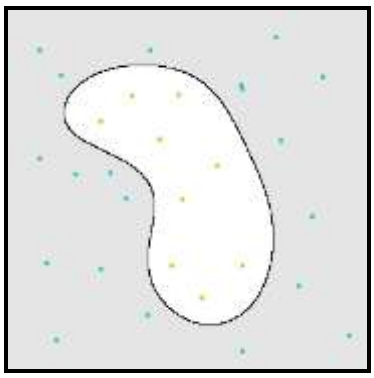

Figure (3): Gaussian Kernel

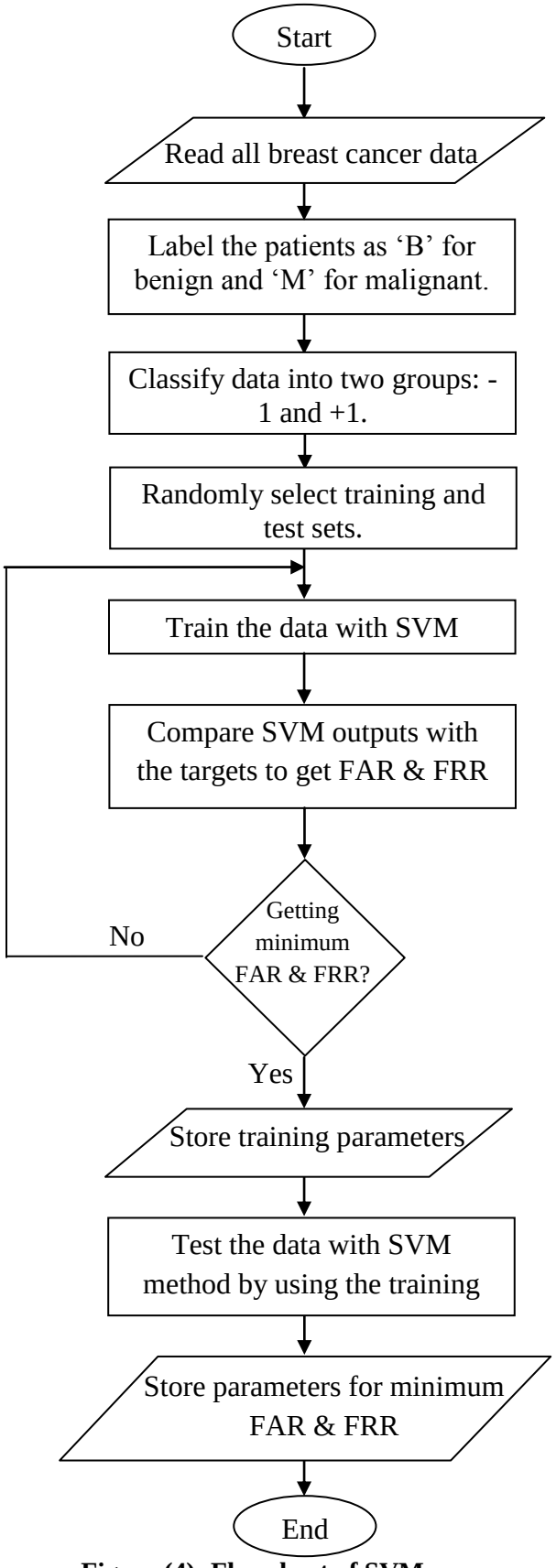

Figure (4): Flowchart of SVM

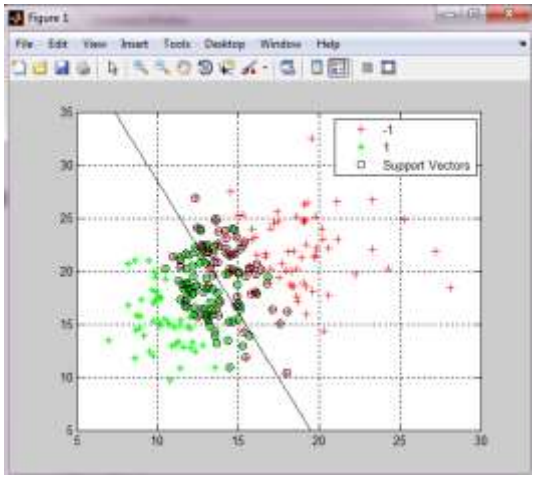

Figure (5): Training result

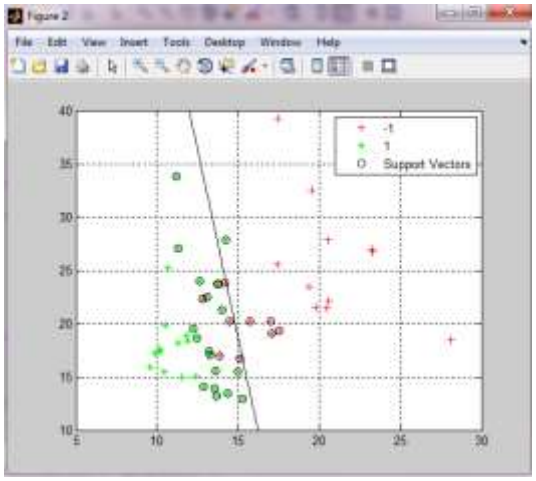

Figure (6): Test result

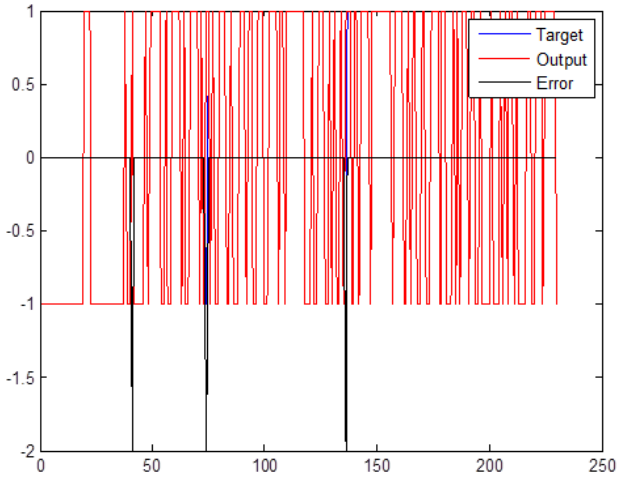

Figure (7): Mean absolute error for train

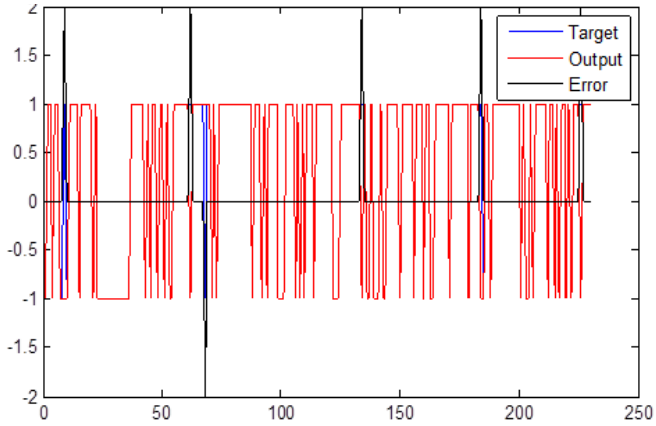

Figure (8): Mean absolute error for test