Automated Classification of Mammographic Abnormalities Using Transductive Semi Supervised Learning Algorithm

Nawel Zemmal[1], Nabiha Azizi[1], Mokhtar Sellami[1], Nilanjan Dey[2]

[1] Labged Laboratory, Computer Science department, Badji Mokhtar University, PO BOX 12, Annaba, 23000. Algeria
[2] Dept of Computer Science, Bengal College of Engineering & Technology, India.
{zemmal, azizi}@labged.net

Abstract. Computer-aided diagnosis (CAD) of breast cancer is becoming a necessity given the exponential growth of performed. CAD are usually characterized by the large volume of acquired data that must be labeled in a specific way that leads to a major problem which is labeling operation. As a result the community of machine learning has attempted to respond to these practical needs by introducing the semi-supervised learning. The motivation of the current research is to propose a TSVM-CAD system for mammography abnormalities detection using a new Transductive TSVM with comparison of its kernel functions. The effectiveness of the system is examined on the Digital Database for Screening Mammography database DDSM using classification accuracy, sensitivity and specificity. Experimental results are very encouraging.

Keywords: Computer Aided Diagnosis (CAD), Transductive Support Vector Machine, Semi Supervised Learning (SSL), mammographic abnormalities.

1 Introduction

The reduction in the death rate caused by this type of cancer as well as favoring the chances of recovery is only possible if the tumor was supported in the early stages of its appearance. Faced with the increasing number of mammograms during the last decades, various researches make the effort to automatically interpret mammogram abnormality through (CAD) systems [1], [2], [3]. In a CAD system, there is in effect always a large volume of data which must be recognized and labeled in a specific way. However, this criterion may not always be satisfied, for reasons of cost or imperfect knowledge of the problem to solve [4]. The statistical learning community has attempted to respond to these practical needs, by formalizing more general problems that supervised learning like semi-supervised learning (SSL) [5].

As the medical images should be represented by different sources of information, it is interesting to integrate different families of characteristics. The used features in this research contain a combination of the three heterogeneous families based on texture and shape which are: co-occurrence matrix, Hu moments and central moments. In the classification step, a semi-supervised classification using Transductive Support Vector

Machine (TSVM) has been opted because of its performance and it is well proven in several areas including the field of medical diagnosis [6], [7].

The remainder of the paper is organized as follows: In Section 2, the semi supervised learning is exposed and the mathematical concept of TSVM is shown. General scheme of proposed approach accompanying with description of each stage is presented in Section 3. Section 4 illustrates the experimental part and the obtained results using TSVM classifier and mixture of features. Conclusion and some perspective points for future extensions achieve the paper.

2 Transductive Support Vector Machine (TSVM)

The SSL classification is achieved, not only with the labeled data, but also with the unlabeled datasets. Among the semi-supervised learning techniques, Transductive Support Vector Machine (TSVM) is the most popular approach and has a strong theoretical base, which inherits from the notion of "large-margin" of supervised SVM [8]. The principle of Transductive SVM algorithm can be expressed as, given a group of independent and identically distributed labeled data sets:

$$D = \{(x1, y1), \dots, (xl, yl)\}, x \in Rn, y \in \{+1, -1\} \quad (1)$$

And unlabeled data sets: $x_1, x_2, \dots, x_k$
The mathematical formula of Transductive SVM can be defined as follows:

$$\text{Min}\left(y1, y2, \dots, yk, \omega, b, \delta 1, \delta 2, \dots, \delta l, \delta_1', \delta_2', \dots, \delta_k'\right) \frac{1}{2}||\omega||^2 + \boldsymbol{C} \sum_{i=1}^{l} \delta_i + \boldsymbol{D} \sum_{i=1}^{k} \delta_j'$$
$$\text{Subject to} \begin{cases} yi(\omega * xi + b) \geq 1 - \delta i;\ \delta i \geq 0; i = 1,2 \dots, l \\ yj(\omega * xj + b) \geq 1 - \delta j;\ \delta_j' \geq 0; j = 1,2 \dots, k \end{cases} \quad (2)$$

where, C, D as a parameter, and D for the impact factor.

3 Proposed CAD System

The proposed CAD represented in Fig. 1 consists into three main steps: tumor contour extraction from the input mammogram images, Features extraction transforms the input data to characteristics and compact representation and classification in order to identify the abnormalities.

Fig. 1. Proposed CAD System

3.1 Tumor Contour Extraction

In the segmentation step, the outline of the shape is extracted and analyzed using a tool for image processing "Image J" [9] and the mammographic image will have the following form (See Fig. 2)

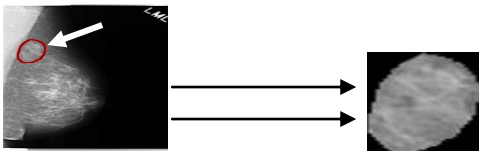

Fig. 2. Extraction of the mass outline by *ImageJ* tool.

3.2 Features Extraction

The methods of image analysis are variable according to the types of features extracted from the image as texture features, the shape features etc.... In the proposed approach, the image is represented by three families of characteristics which are: the co-occurrence matrix [10], Hu moments [11] and central moments [12]. For more detail of these three families, the reader is referred to the previous study in [13].

3.3 Learning and Classification

Transductive SVM is chosen in this study because of its performance in several areas such as medical diagnosis comparing with other classification techniques.

In this work, Digital Database for Screening Mammography (DDSM) is used as test dataset [14]. It was assembled by a group of researchers from University of South Florida. The DDSM database contains 2620 cases.

4 Results and Discussion

The proposed system for abnormalities breast diagnosis has been trained on a sample of 400 images (benign and malignant) taken from DDSM data base. Proposed system was built using JAVA SE 1.8.0 with a simple user interface.

4.1 Classification Performance

Proposed TSVM-CAD for mammogram abnormalities detection and classification using a sample of 200 images (90 malignant and 110 benign) from the DDSM dataset. After contour extraction of the mass and feature extraction step (co-occurrence matrix, Hu moments and central moments). This feature vector contains over than 26 characteristics. Thereafter, a Transductive Support Vector Machine (TSVM) classifier

is applied. During this phase, several empirical tests are made to keep the ones that generate the highest rate of classification. First, each family was tested independently with the Transductive SVM classifier then all features are grouped together in a single vector that will be the entrance of the classifier. Also, three best-known kernel functions are tested (Gaussian, Triangle and Linear). Table 1 summarizes the accuracy rates of each kernel function with the three families of characteristics.

Table 1. Obtained Results of TSVM classifier with the three families of characteristics and the different kernel functions.

Kernel Functions	TSVM with co-occurrence Matrix	TSVM with Hu Moments	TSVM with central moments	TSVM with all features
Gaussian	84,13%	80,10%	73,58%	92,95%
Triangle	77,65%	72,25%	69,95%	89,44%
Linear	75,58%	70,13%	67,13%	82,11%

In order to evaluate the classification performance, other related metrics are also calculated as (see Table 2):

$$\textbf{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\textbf{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{5}$$

Where TP (resp. TN) is True Positive (resp. true Negative) and FP (resp. FN) is False Positive (resp. False Negative).

Table 2. Obtained results of the different metrics of the proposed CAD System.

Kernel Functions	**Sensitivity**	**Specificity**	**Accuracy**
Gaussian	0,89	0,93	0,929
Triangle	0,82	0,86	0,894
Linear	0,79	0,82	0,821

From this table TSVM has again proved its effectiveness in the field of medical diagnosis and especially in indentifying the abnormalities in the mammographic images with a high accuracy rate (92, 95%).

6 Conclusion

Current study was conducted to determine the right path for the future evolution of image processing in medicine and health. In this work, a new tool for the diagnosis of mammography abnormalities is proposed which is based on semi-supervised classification.

During the learning phase TSVM, several empirical tests are made to keep the ones that generate the highest rate of classification. Also, the three best-known kernel functions (Gaussian, Triangle and Linear) are tested in order to analyze the behavior of the classifier TSVM. However, Transductive SVM takes all unlabeled data without prior selection of data that allows better accuracy what can be considered as a drawback. As future study, it will be interesting to analyze how to overcome the latest inconvenient by introducing new concepts to increase the performance of Transductive SVM classifier and to select the most appropriate unlabeled images using active learning as obvious solution.

References

1. Cedolini, C., Bertozzi, S., Londero, A., Bernardi, S., Seriau, L., Concina, S.: Type of Breast Cancer Diagnosis, Screening, and Survival. Clinical Breast Cancer, Vol. 14 (2014) 235-240.
2. Zheng, B., Yoon, S., Lam, S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, Vol. 41 (2014) 1476-1482.
3. Abbadi, N.E., Taee, E.A.: Breast Cancer Diagnosis by CAD, International Journal of Computer Applications, Vol. 100 (2014) 25-29.
4. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recognition, Vol. 42 (2009) 334 – 348.
5. Chapelle, O., Scholkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press, Cambridge, MA, (2006)
6. Zadeh, H.S., Rad, F., Pourabdollah, N.: Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms. Pattern Recognition, Vol. 37 (2004) 1973-1986.
6. Filipovych, R., Davatzikos, C.: Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). NeuroImage, Vol. 55 (2011) 1109-1119.
7. Pang, S., Ban, T., Kadobayashi, Y., Kasabov, N.: Personalized mode transductive spanning SVM classification tree. Information Sciences, Vol. 181 (2011) 2071–2085.
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag (1996).
9. Grishagin, I.V.: Automatic cell counting with ImageJ. Analytical Biochemistry [In Press].
10. Nanni, L., Brahnam, S., Ghidoni, S., Menegatti, E.: A comparison of methods for extracting information from the co-occurrence matrix for subcellular classification. Expert Systems with Applications, Vol. 41 (2013) 7457-7467.
11. Žunić, D., Žunić, J.: Shape ellipticity from Hu moment invariants. Applied Mathematics and Computation, Vol. 226 (2014) 406-414.
12. Grubbström, D., Tang, O.: The moments and central moments of a compound distribution. European Journal of Operational Research, Vol. 170 (2006) 106-119.
13. Azizi, N., Zemmal, N., Tlili, Y.: Kernel Based classifiers fusion with features diversity for breast masses classification. 8th international Workshop on Systems, Signal Processing and their Application (WoSSPA) (2013) 116-121.
14. Heath, M.D., Bowyer, K.W.: Mass detection by relative image intensity. In 5th International Workshop on Digital Mammography, Toronto, Canada (2000) 219–225.