Artificial Intelligence in Medicine 65 (2015) 219–227

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification

Bartosz Krawczyk[a,*], Gerald Schaefer[b], Michał Woźniak[a]

[a] Department of Systems and Computer Networks, Wrocław University of Technology, Wyb. Wyspianskiego 27, 50-370 Wrocław, Poland
[b] Department of Computer Science, Loughborough University, Loughborough LE11 3TU, UK

ARTICLE INFO

Article history:
Received 11 May 2013
Received in revised form 15 July 2015
Accepted 23 July 2015

Keywords:
Classifier ensemble
Imbalanced classification
Cost-sensitive classification
Ensemble pruning
Evolutionary algorithm
Breast cancer detection
Thermogram

ABSTRACT

Objectives: Early recognition of breast cancer, the most commonly diagnosed form of cancer in women, is of crucial importance, given that it leads to significantly improved chances of survival. Medical thermography, which uses an infrared camera for thermal imaging, has been demonstrated as a particularly useful technique for early diagnosis, because it detects smaller tumors than the standard modality of mammography.
Methods and material: In this paper, we analyse breast thermograms by extracting features describing bilateral symmetries between the two breast areas, and present a classification system for decision making. Clearly, the costs associated with missing a cancer case are much higher than those for mislabelling a benign case. At the same time, datasets contain significantly fewer malignant cases than benign ones. Standard classification approaches fail to consider either of these aspects. In this paper, we introduce a hybrid cost-sensitive classifier ensemble to address this challenging problem. Our approach entails a pool of cost-sensitive decision trees which assign a higher misclassification cost to the malignant class, thereby boosting its recognition rate. A genetic algorithm is employed for simultaneous feature selection and classifier fusion. As an optimisation criterion, we use a combination of misclassification cost and diversity to achieve both a high sensitivity and a heterogeneous ensemble. Furthermore, we prune our ensemble by discarding classifiers that contribute minimally to the decision making.
Results: For a challenging dataset of about 150 thermograms, our approach achieves an excellent sensitivity of 83.10%, while maintaining a high specificity of 89.44%. This not only signifies improved recognition of malignant cases, it also statistically outperforms other state-of-the-art algorithms designed for imbalanced classification, and hence provides an effective approach for analysing breast thermograms.
Conclusions: Our proposed hybrid cost-sensitive ensemble can facilitate a highly accurate early diagnostic of breast cancer based on thermogram features. It overcomes the difficulties posed by the imbalanced distribution of patients in the two analysed groups.

1. Introduction

Breast cancer is the most commonly diagnosed form of cancer in women, accounting for about 30% of all cases [1]. Medical thermography, which uses a thermal-imaging infrared camera to capture a temperature distribution of the human body, has been shown to be well-suited for the task of detecting breast cancer [2,3]. In contrast to other modalities such as mammography, thermography is a non-invasive, non-contact, passive, and radiation-free technique. The radiance from human skin is a function of the surface temperature which in turn is influenced by the level of blood perfusion in the skin. Thermal imaging is hence capable of detecting changes in blood perfusion which may be the result of inflammation, angiogenesis, or other causes [4]. Asymmetrical temperature distributions as well as hot or cold spots are known to be strong indicators of an underlying dysfunction [5].

It has been shown that thermography has advantages over the standard modality of mammography for breast cancer diagnosis, in particular when the tumor is in its early stages or in dense tissue. Early detection is crucial as it provides significantly higher chances of survival [6], and in this respect infrared imaging can outperform mammography. While mammography can detect tumors only after they exceed a certain size, even small tumors can be identified using thermography, because the high metabolic activity of cancer cells leads to an increase in local temperature detectable by

* Corresponding author. Tel.: +48 692979578.
E-mail addresses: bartosz.krawczyk@pwr.edu.pl (B. Krawczyk), gerald.schaefer@ieee.org (G. Schaefer), michal.wozniak@pwr.edu.pl (M. Woźniak).

http://dx.doi.org/10.1016/j.artmed.2015.07.005
0933-3657/

infrared imaging. Published analyses report that the average tumor size undetected by mammography is 1.66 cm compared to only 1.28 cm by thermography [7].

In this paper, we present an effective approach for analysing thermographic breast images to detect breast cancer. The basis of our approach is the characterisation of image features describing the bilateral symmetry between the two breast regions, because tumors lead to asymmetries between the temperature distributions of the two sides. The image features we employ include basic statistical features, histogram features, image moments, and various texture features.

The derived image features are then used during a pattern classification stage. While numerous classification algorithms have been published [8], it is well known that no method is superior for all possible decision problems [9]. Consequently, it is common to train several predictors and then select a model that displays desirable properties for the problem at hand. An alternative is to employ a multiple classifier system, or ensemble classifier, considers the decisions of a committee of individual classifiers [10].

There are several factors that may strongly affect the performance of classification algorithms. An important aspect, in particular in the medical domain [11], is the underlying class distribution. Most learning models assume that the distribution of samples is roughly equal between the classes. An imbalanced binary classification problem [12] occurs when the number of samples in one class (the majority class) significantly outnumbers that in the other class (the minority class). Imbalanced datasets frequently appear in a variety of applications including medical data analysis [13], anomaly detection [14], handwritten symbol recognition [15], or object detection [16]. While classifier performance is typically evaluated based on predictive accuracy, this is not appropriate for imbalanced data as it would favour a bias towards the majority class.

Furthermore, and again particularly in medical decision making, it is often the minority class that is of higher importance. Clearly, the costs associated with missing a cancer case (false negative) are much higher than those for mislabelling a benign one (false positive). Therefore, standard classifiers aimed only at maximising overall classification accuracy have no means of considering this, inevitably resulting in an inferior decision making system [17].

In this paper, we address these challenges and propose, based on our earlier work [18], a hybrid cost-sensitive classifier ensemble algorithm for effective breast thermogram feature analysis. We use cost-sensitive decision trees as base classifiers, because they are readily improved with the ensemble approach. Instead of using a pre-defined cost matrix, we derive its parameters through receiver operating characteristic (ROC) analysis. To generate our ensemble, we employ an evolutionary algorithm to simultaneously perform feature selection and classifier fusion. In that context, we propose a combination of cost and diversity criteria to form a pool of mutually complementary classifiers which provides excellent sensitivity. We furthermore prune the ensemble by removing predictors that contribute minimally to the overall decision, thereby reducing the computational complexity of the ensemble. Extensive experimental results demonstrate that our approach achieves improved recognition of cancer cases, statistically outperforming several state-of-the-art algorithms designed for imbalanced classification. The main contributions of this paper are as follows:

- A new hybrid ensemble approach based on combining cost-sensitive decision trees for efficient breast thermogram classification.
- Use of an evolutionary algorithm for simultaneous feature selection and weight assignment for classifier fusion. This method explores diverse subsets of features and promotes the best individual classifiers to boost the recognition rate of the malignant class.
- A combined criterion for the training algorithm which simultaneously minimises the overall misclassification cost and assures diversity among the pool of classifiers.
- Analysis of the influence of the cost matrix parameters (derived based on ROC analysis) on the performance of the proposed ensemble.
- A pruning procedure for discarding classifiers with low influence on the final decision, decreasing the computational complexity of the ensemble.

The remainder of the paper is organised as follows. In Section 2 we describe the analysed breast thermogram data, while Section 3 discusses the problem of imbalanced classification. Our new algorithm is introduced in detail in Section 4. Experimental results are reported and discussed in Section 5, while Section 6 concludes the paper.

2. Breast thermogram feature analysis

While both frontal and/or lateral-view thermograms can be imaged for breast cancer diagnosis, in this paper, we restrict out attention to frontal-view images. As has been shown [19], analyzing the symmetry between the thermal images of the left and right breast is an effective approach to automatically detecting cancer cases. A malignant tumor recruits blood vessels resulting in hot spots and a change in the vascular pattern. This causes an asymmetry between the temperature distributions of the two breasts [20]. In contrast, symmetrical thermal images typically signify the absence of breast cancer.

We segment the areas corresponding to the left and right breast from the thermograms. Once segmented, we convert the breast regions to a polar coordinate representation to simplify the calculation of several of our features. A series of statistical features is then calculated to provide indications of symmetry between the regions of interest (i.e., the two breast areas) [21].

The simplest feature describing the temperature distribution captured in thermograms is the statistical mean. Since we are interested in symmetry features, we calculate the mean for each breast and use the absolute difference between the two. Similarly, we calculate the standard temperature deviation and use the absolute difference as a feature. Furthermore, we employ the absolute differences of the median temperature and the 90th percentile temperature.

Image moments [22] are defined as

$$m_{pq} = \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} x^p y^q g(x,y), \tag{1}$$

where x and y denote the pixel location and N and M the image size. We utilise moments m_{01} and m_{10} which essentially describe the centre of gravity of the breast regions, as well as the distance (in both the x and y directions) of the centre of gravity from the geometrical centre of the breast. For all four features, we calculate the absolute differences of the values between the left and right breast.

Histograms record the frequencies of certain temperature ranges in the thermograms. In our work, we construct normalised histograms for each breast and use the cross-correlation between the two histograms as a feature. From the difference histogram (i.e., the difference between the two histograms), we compute the absolute value of its maximum, the number of bins exceeding a certain threshold (0.01 in our experiments), the number of zero crossings, energy and the difference of the positive and negative parts of the histogram.