Applied Soft Computing 38 (2016) 714–726

Contents lists available at ScienceDirect

Applied Soft Computing

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

Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy

Bartosz Krawczyk[a,*], Mikel Galar[b], Łukasz Jeleń[c], Francisco Herrera[d,e]

[a] Department of Systems and Computer Networks, Wrocław University of Technology, Wrocław, Poland
[b] Departamento de Automática y Computación, Universidad Pública de Navarra, Pamplona, Spain
[c] Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wrocław, Poland
[d] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[e] Faculty of Computing and Information Technology – North Jeddah, King Abdulaziz University, 21589 Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 24 April 2015
Received in revised form 23 July 2015
Accepted 24 August 2015
Available online 9 October 2015

Keywords:
Machine Learning
Classifier ensemble
Imbalanced classification
Evolutionary algorithms
Clinical decision support
Breast cancer

ABSTRACT

In this paper, we propose a complete, fully automatic and efficient clinical decision support system for breast cancer malignancy grading. The estimation of the level of a cancer malignancy is important to assess the degree of its progress and to elaborate a personalized therapy. Our system makes use of both Image Processing and Machine Learning techniques to perform the analysis of biopsy slides. Three different image segmentation methods (fuzzy c-means color segmentation, level set active contours technique and grey-level quantization method) are considered to extract the features used by the proposed classification system. In this classification problem, the highest malignancy grade is the most important to be detected early even though it occurs in the lowest number of cases, and hence the malignancy grading is an imbalanced classification problem. In order to overcome this difficulty, we propose the usage of an efficient ensemble classifier named EUSBoost, which combines a boosting scheme with evolutionary undersampling for producing balanced training sets for each one of the base classifiers in the final ensemble. The usage of the evolutionary approach allows us to select the most significant samples for the classifier learning step (in terms of accuracy and a new diversity term included in the fitness function), thus alleviating the problems produced by the imbalanced scenario in a guided and effective way. Experiments, carried on a large dataset collected by the authors, confirm the high efficiency of the proposed system, shows that level set active contours technique leads to an extraction of features with the highest discriminative power, and prove that EUSBoost is able to outperform state-of-the-art ensemble classifiers in a real-life imbalanced medical problem.

1. Introduction

Based on the data provided by the National Cancer Registry, up to 2015 there were 17,144 diagnosed cases of breast cancer in Poland. This statistic makes the breast cancer the most often diagnosed type of cancer among middle-age women and the number of diagnosed cases is still increasing. For instance, between 2009 and 2012 there was an increase of 1280 diagnosed cases. Unfortunately, this fact is translated into a larger death rate, which was recorded to be 5651 deaths in 2012, 341 more cases than in 2009. However, most of them could have been fully recovered if the diagnosis would have been made in the early stage of the disease. This is because cancers in their early stages are vulnerable to treatment, while cancers in their most advanced stages are usually almost impossible to treat.

In order to differentiate the stages of a cancer, during the diagnosis process a grade is assigned, which is then used to determine the appropriate treatment. Since successful treatment is a key to reduce the high death rate of breast cancer, so it is the appropriate grading of the cancer malignancy. For this purpose, screening mammography tests are performed and when a suspicious region is found a fine needle aspiration biopsy (FNA) is taken. This is an invasive method, which extracts a small sample of the questionable breast tissue that allows the pathologist to describe the type of the cancer in detail. Malignancy grading allows doctors to precisely estimate cancer behavior with or without undertaking treatment,

* Corresponding author. Tel.: +48 692979578.
E-mail addresses: bartosz.krawczyk@pwr.edu.pl (B. Krawczyk), mikel.galar@unavarra.es (M. Galar), lukasz.jelen@pwr.edu.pl (Ł. Jeleń), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.asoc.2015.08.060
1568-4946/

and therefore is called a prognostic factor. It plays an important role in breast cancer diagnosis and the appropriate treatment is chosen accordingly to this factor.

The determination of malignancy is performed by assigning a malignancy grade to the case. To help in this very difficult task, a grading scale was proposed by Bloom and Richardson [6]. The grading scheme proposed by the authors was derived to assess malignancy from histological slides and is now widely used among pathologists to grade not only histological but also cytological tissues. However, due to the large variation in cancer images and the large number of slides to be analyzed every day by a specialist, a need for automatic decision support system has arisen. Machine Learning is a popular tool for developing support software that ease the work of the specialists. Among a plethora of methods used in this domain, classifier ensembles stand as one of the most efficient solutions for image classification [15,39].

Automatic breast cancer detection from medical images has been widely addressed in the contemporary literature [14,48]. There are numerous reports on applying different imaging techniques (such as microscopic analysis [19], mammography [61] or magnetic resonance [49]), segmentation methods [38] or classification approaches [20] for this task. However, not much attention was paid to the problem of designing a decision support system for breast cancer malignancy grading [36].

Classification of malignancy grading suffers from a well-known difficulty in Machine Learning, the class imbalance problem [29,22]. This problems arises when one class appears much more often than the other (we have many more cases from medium malignancy that from high malignancy), which leads to an uneven distribution of examples in the training set. This is a challenging problem in Machine Learning [67], since it usually brings along a number of difficulties such as overlapping, small disjuncts and small sample size [46]. For these reasons, specific methods to address these types of problems are needed.

Imbalanced classification must be carefully addressed in the context of breast cancer classification. Standard classification methods tend to get biased towards the majority class, ignoring the minority class or treating it as noise. However, the minority class is the most important one – in the discussed application it corresponds to the highest grade of malignancy posing a significant threat to the life of a patient. Therefore, it must be detected with the highest precision, as one cannot allow for such a severe case to be mistreated.

In this paper, we discuss the application of Pattern Recognition and Image Processing methods to extract the information from the FNA slides and automatically assign a malignancy grade to the case. In order to do so, we consider three different methods for segmenting cytological images and extracting features from them. Then, we apply a highly efficient ensemble fitted for handling difficult imbalanced problems. This method is based on a combination of boosting method [21] with evolutionary undersampling [26]. This allows us to propose a complete, automatic and highly efficient clinical decision support system that can be used in a daily physician routine.

The main contributions of this work are as follows:

- An automatic and complete clinical decision support system for breast cancer malignancy grading is developed.
- Different methodologies for segmenting fine needle aspiration biopsy (FNA) slides and extracting meaningful features from them are examined.
- An efficient classifier ensemble, specifically designed for imbalanced problems with difficult data distribution is considered. Boosting scheme is combined with evolutionary undersampling to obtain both accurate and diverse base classifiers.
- An extensive experimental analysis is carried out on a large database collected by the authors, showing that the proposed evolutionary undersampling boosting can outperform state-of-the-art methods dedicated to binary imbalanced learning, and hence proving the usefulness of the designed approach to breast cancer malignancy grading.

The remaining part of this paper is organized as follows. Section 2 gives essential background about the problem of breast cancer malignancy grading. In Section 3, we discuss the three algorithms used to segment FNA slides. Section 4 introduces the imbalanced classification domain and reviews current algorithms and measures used in this field. The proposed evolutionary undersampling boosting ensemble is presented in detail in Section 5. In Section 6 the set-up used for the experimental analysis (methods used and their parameters), the results obtained and the discussion can be found, whereas Section 7 concludes the paper.

2. Breast cancer malignancy grading

Malignancy grading is one of the most important steps during cancer diagnosis. Based on that grading doctors are able to determine the appropriate treatment and, what is even more important, predict if the undertaken treatment is going to be successful. This examination is performed when suspicious regions in the breast tissue are found. For this purpose a mammography examination is done. When a suspicious region is found, a fine needle aspiration biopsy is taken. This is an invasive procedure that involves the extraction of a breast tissue with a syringe with outer needle diameter smaller than 1 mm (typically between 0.4 and 0.7 mm). The tissue is then place on a glass slide, stained and examined under a light microscope.

This examination is based on the well defined scheme given by Bloom and Richardson [6] and called accordingly the Bloom–Richardson grading scheme. This scheme has underwent many modifications and currently the modification proposed by Scarff, called a modified Scarff–Bloom–Richardson system, is used to grade the breast cancer malignancy [55]. The grading scheme describes several features that are divided into three groups known as factors that assesses features in a point based scale.

1. *Degree of structural differentiation (SD).* This factor describes cells' ability to form groups. In histopathological slides, for which the grading scheme is also used, this factor describes cell tendency to form tubules. In the cytological smears tubules are not preserved and therefore cells' groupings are examined. This factor is visualized in Fig. 1, where intermediate malignancy case with one group (Fig. 1a) and high malignancy case with highly dispersed cells (Fig. 1b) are presented.
2. *Pleomorphism (P).* This factor examines differences in size, shape and staining of the nuclei[1]. This scoring is fairly straightforward because the greater the irregularity of the nuclei is, the worse the prognosis becomes, as it can be observed in Fig. 2a.
3. *Frequency of hyperchromatic and mitotic figures (HMF).* This factor assesses the number of visible mitosis in the image. Mitosis is the process in the cell life cycle in which a mother cell is divided into two identical cells. From the Image Processing point of view, mitosis can be observed as a dark stain in the nucleus. Here, the more mitotic cells are, the worse the prognosis is. An example of the mitosis is shown in Fig. 2b.

[1] Nucleus is a a central organelle of a cell that contains most of the cell's DNA.