Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy

Article *in* Applied Soft Computing · October 2015

DOI: 10.1016/j.asoc.2015.08.060

4 authors:

Bartosz Krawczyk
Virginia Commonwealth University
123 PUBLICATIONS **494** CITATIONS
SEE PROFILE

Mikel Galar
Universidad Pública de Navarra
63 PUBLICATIONS **954** CITATIONS
SEE PROFILE

Lukasz Jelen
Wroclaw University of Science and Technology
32 PUBLICATIONS **84** CITATIONS
SEE PROFILE

Francisco Herrera
University of Granada
872 PUBLICATIONS **37,740** CITATIONS
SEE PROFILE

Applied Soft Computing xxx (2015) xxx–xxx

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 xxx

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

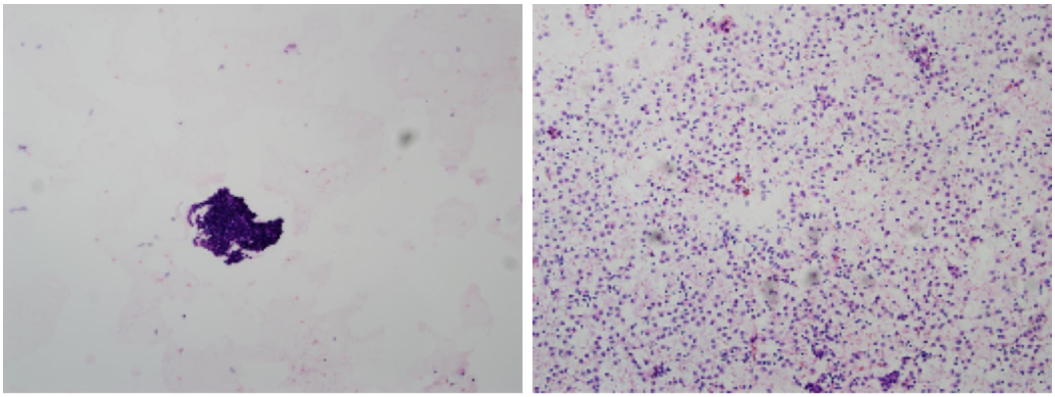
Malignancy grading is one of the most important steps during cancer diagnosis. Based on that grading doctors are able to determine the appropriate treatment and, what is even more important, predict if the undertaken treatment is going to be successful. This examination is performed when suspicious regions 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.

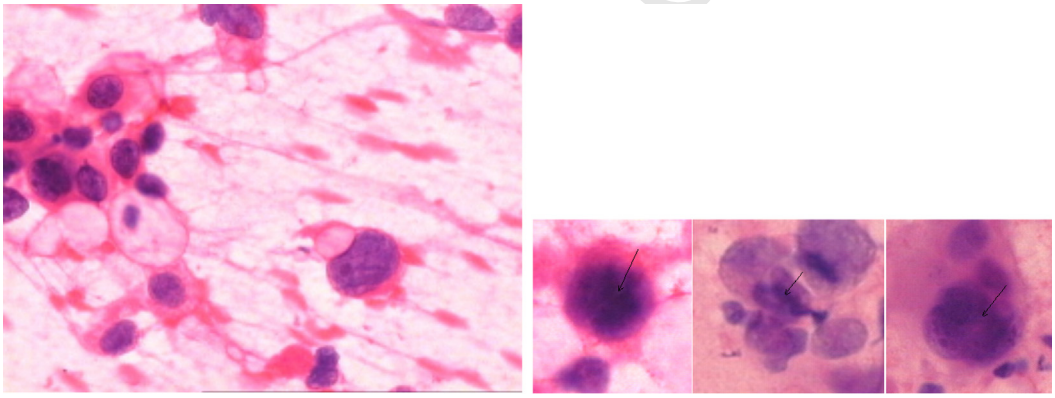
[1] Nucleus is a a central organelle of a cell that contains most of the cell's DNA.

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060



(a) Intermediate malignancy case (b) High malignancy case

Fig. 1. Illustration of the structural differentiation. (a) Intermediate malignancy case. (b) High malignancy case.



(a) Pleomorphism (b) Mitosis [35]

Fig. 2. Illustration of the Pleomorphism features and mitosis. (a) Pleomorphism. (b) Mitosis [35].

All three factors of the modified Scarff–Bloom–Richardson scheme are assessed in a three-point scale, where one point is assigned to the least malignant case and three points are given to the highest malignancy one. Furthermore, the final grade is obtained summing up the quantitative values assigned to the three features (Eq. (1)) and following the chart presented in Fig. 3.

$$G = SD + P + HMF. \quad (1)$$

Nonetheless, even though these factors are well-explained, the determination of cancer malignancy is a very difficult task and depends not only on the experience of the pathologist but also on his/her mind. More experienced pathologists that have seen more cases are more reliable in their diagnosis. However, due to overwork and fatigue, seeing more similar cases may lead to misclassification of the malignancy grade. In order to address this problem we present an automated grading approach that is able to evaluate and assign a grade to a FNA biopsy tissue, that is, we translate the modified Scarff–Bloom–Richardson grading scheme into a classification problem.

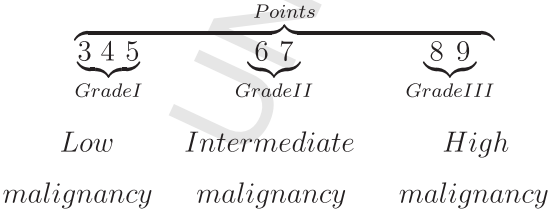


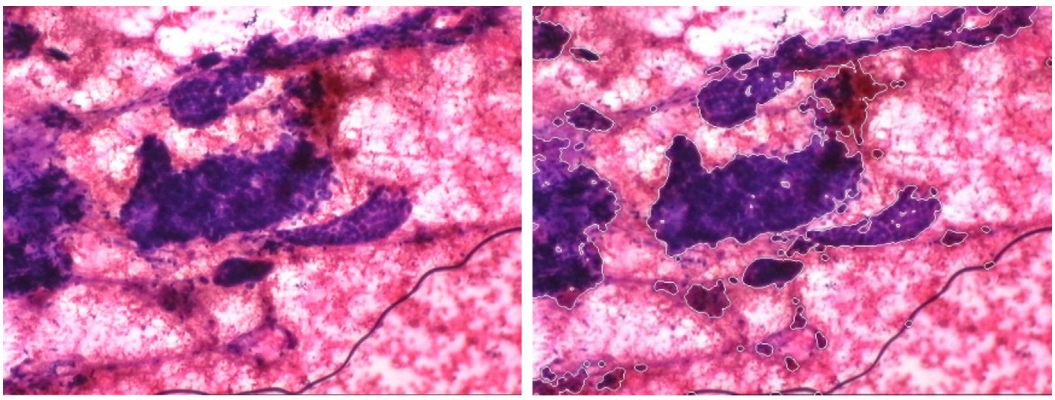
Fig. 3. Grade determination for the Scarff – Bloom – Richardson scheme [6].

3. Image segmentation

One of the objectives of this paper is to perform a comparative study of three segmentation techniques to determine which one is the best method for nuclei extraction with respect to the posterior best breast cancer malignancy classification process. In computer vision, segmentation is a very important task because it influences all the subsequent stages such as feature extraction and classification. This is why it is extremely important to find a technique capable of extracting the nuclei accurately.

In the proposed framework, the segmentation of breast cancer cells and nuclei is divided into two parts. In the first part we segment a low magnification image, whereas in the second a high magnification image is segmented (please refer to the dataset description in Section 6.1 for more details about the type of images). According to the modified Scarff–Bloom–Richardson scheme, for the 100× magnification images, we only need information about cells' groupings. This is because we are interested in structure of cells' groups, not in their individual shapes. We have found that for this purpose, a simple automatic thresholding method is sufficient to achieve an accurate segmentation. The threshold level was determined based on the bimodal histogram of the image with the algorithm described by Riddler and Calvard [53]. The segmentation results are shown in the Fig. 4.

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060



(a) Original image (b) Segmented image

Fig. 4. Segmentation of the 100× magnification case. (a) Original image. (b) Segmented image.

Otherwise, for the high magnification images, a more advanced segmentation algorithm needs to be applied. This is due to the fact that these images are used for the determination of cellular features, which need to be as accurate as possible. This is why a precise representation of the nucleus is required. For this purpose, we carry out a comparison of three segmentation techniques: fuzzy c-means color segmentation [37], level set active contours technique [47] and grey-level quantization method that uses a texture information for segmentation [28].

3.1. Fuzzy c-means

The first segmentation method implemented in this study is the fuzzy c-means approach (FCM) proposed by Klir and Yuan [37]. It is focused on dividing the data $X = \{x_1, x_2, \ldots, x_n\}$ into c clusters assuming that there is a known pseudo-partition $P = \{A_1, A_2, \ldots, A_c\}$ and A_i is a vector of all memberships of x_k to cluster i. Applying Eq. (2) we can calculate centers of the c clusters [60].

$$v_i = \frac{\sum_{k=1}^{n} [A_i(x_k)]^m x_k}{\sum_{k=1}^{n} [A_i(x_k)]^m}, \qquad i = 1, 2, \ldots, c, \tag{2}$$

where $m > 1$ is a weight controlling the fuzzy membership. If $\|x_k - v_i\|^2 > 0$ for all $i \in \{1, 2, \ldots, c\}$ then the memberships are defined by Eq. (3). If $\|x_k - v_i\|^2 = 0$ for some $i \in I \subseteq \{1, 2, \ldots, c\}$ the memberships are defined as a nonnegative real number satisfying Eq. (3) for $i \in I$.

$$A_i(x_k) = \left[\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{3}$$

where $\sum_{i \in I} A_i(x_k) = 1$.

We have noticed that values of m between 2 and 4 did not have significant influence on the segmented nuclei. Therefore the value of $m = 2$ was used for segmentation.

In order to segment an image we look for a set P that minimizes the performance index $J_m(P)$ (Eq. (4)). The optimization solution to this problem can be found in [4].

$$J_m(P) = \sum_{k=1}^{n} \sum_{i=1}^{c} [A_i(x_k)]^m \|x_k - v_i\|^2. \tag{4}$$

3.2. Level sets

The clustering method based on level sets (LS) is the second segmentation algorithm considered to extract precise nuclear information. Level sets belong to active contours models because they change their shape according to the information in the image. This property makes them a very good choice in biomedical applications [47].

Level sets were introduced by Osher and Sethian [50] as a method for capturing moving fronts. This method relays on the determination of a surface $\Gamma(t)$ described by Eq. (5). The surface propagates along its normal direction and is embedded as a zero level of a time–varying higher dimensional function $\phi(x, t)$ [50].

$$\Gamma(t) = \{x \in \mathcal{R}^3 / \phi(x, t) = 0\}. \tag{5}$$

The determination of Γ, which is a closed curve in R^2, requires a definition of an evolution equation for ϕ. This can be expressed in general form as [57]:

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0. \tag{6}$$

The function ϕ describes a curve defined by $\phi(x, t) = d$, where d is a signed distance between x and the surface Γ. If x is inside (outside) of Γ then d is negative (positive). Function F is a scalar speed function that depends on image data and the function ϕ.

Here ϕ is initialized as a signed distance function before evolution and is periodically reshaped to be a signed distance function. This is due to the fact that during the evolution, ϕ can assume sharp or flat shapes [43].

In this study we have applied a modified level set approach of Li et al. [43] that overcomes the ϕ reshaping problem according to the evolution equation of a form:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial \mathcal{E}}{\partial \phi}, \tag{7}$$

where $\frac{\partial \mathcal{E}}{\partial \phi}$ is a Gateaux derivative of the energy function $\mathcal{E}$ and is represented by:

$$\frac{\partial \mathcal{E}}{\partial \phi} = -\mu \left[\Delta \phi - div \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right] - \lambda \delta(\phi) div \left(g \frac{\nabla \phi}{|\nabla \phi|} \right) - \nu g \delta(\phi), \tag{8}$$

where Δ is the Laplacian operator, div is the divergence operator, $\mu > 0$ is a parameter controlling the effect of penalizing the deviation of ϕ from a signed distance function, g is an edge indicator function, $\lambda > 0$ and ν are constants. In this study we have used $\mu = 0.04$, $\lambda = 5.0$ and $\nu = 1.0$.

3.3. Grey-level quantization

The third segmentation technique is a grey-level quantization (GLQ) based on the image textural description. It uses a second order statistic for the generation of grey level co-occurrence texture features [28]. In this method, for a spatial window inside the image, the conditional joint probabilities C_{ij} are calculated according to Eq.

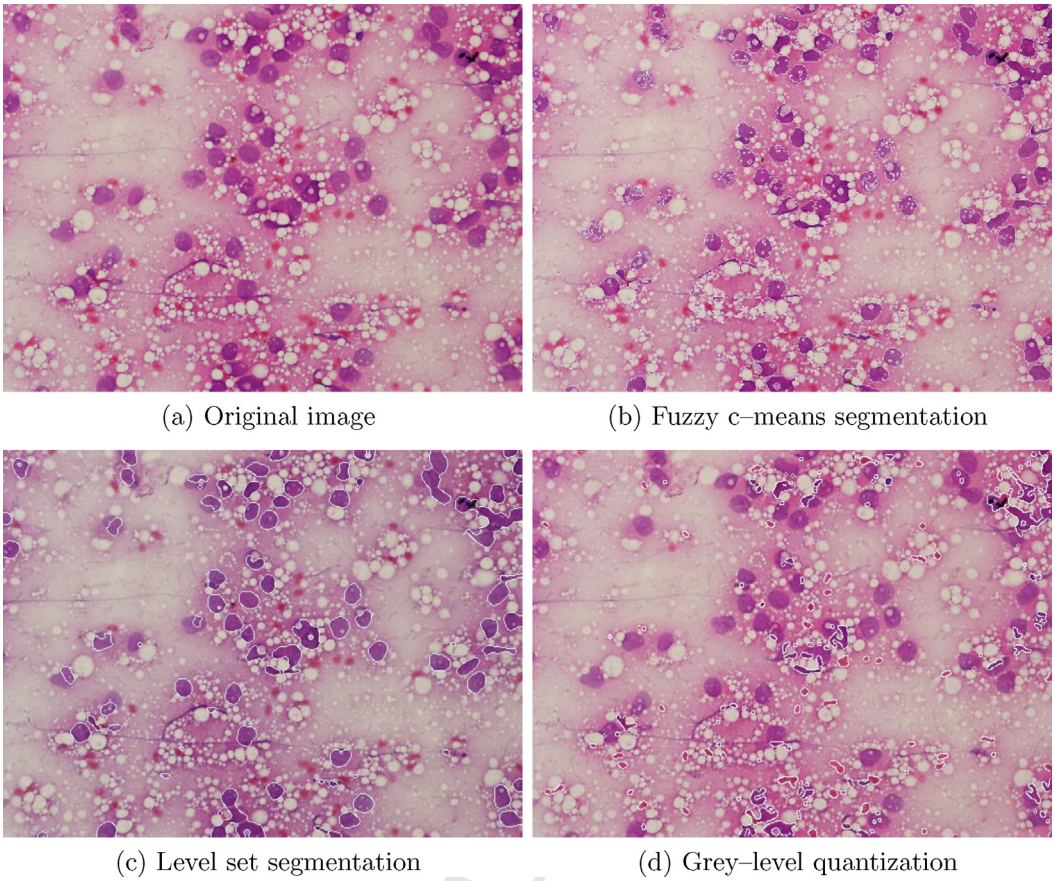


Fig. 5. Segmentation results for high magnification images. (a) Original image. (b) Fuzzy c-means segmentation. (c) Level set segmentation. (d) Grey-level quantization.

(9) for all pairwise combinations of grey levels. We assume that the distance between the pixels is known.

$$C_{ij} = \frac{P_{ij}}{\sum_{i,j=0}^{G-1} P_{ij}}, \tag{9}$$

where P_{ij} is a frequency of occurrence of two grey levels i and j and G is the number of quantized grey levels.

The probabilities are stored as a gray level co-occurrence matrix, where the (i, j) element of the matrix represents the probability C_{ij}. To identify textures within an image four features from the dependency matrix were derived. These properties are described by the following equations:

$$Entropy = -\sum_{i,j=0}^{G-1} C_{ij} \ln C_{ij}, \tag{10}$$

$$Contrast = \sum_{i,j=0}^{G-1} C_{ij}(i-j)^2, \tag{11}$$

$$Inertia = \sum_{i,j=0}^{G-1} \frac{(i-\mu_x)(j-\mu_y)C_{ij}}{\sigma_x \sigma_y}, \tag{12}$$

$$Energy = \sum_{i,j=0}^{G-1} C_{ij}^2, \tag{13}$$

where σ is the standard deviation and μ is the mean.

3.4. Segmentation results

To assess the segmentation results, all the images were visually studied by an expert pathologist to check which algorithm provided the best results. In Fig. 5, an example of the results of all three segmentation algorithms are presented. It can be noticed that level sets method provides the most accurate boundary representation. The main drawback of this method is that it requires an initial boundary representation that is later reshaped (initial level set). To make the proposed framework fully automatic we have obtained the initial level set as a result of the automatic thresholding method. Both fuzzy c-means and grey-level quantization are able to efficiently segment nuclei without any prior information about the boundary. From Fig. 5 one can observe that although FCM provides better nuclei segmentation than GLQ, the level sets based method is the one with the best representation of the nuclear boundary, whereas GLQ algorithm loses a lot of nuclear information during the segmentation. In Section 6, we will show the influence of this observations on the classification of the breast cancer malignancy.

4. Imbalanced classification

As we have already mentioned, the breast cancer malignancy grading classification considered in this study is an imbalanced classification problem, where the number of instances from one class is greater than the number of instances of the other class. As we describe in Section 6.1, there are 137 examples of class G2 (intermediate malignancy) and 39 of class G3 (high malignancy), whereas there are no cases of low malignancy. Hence, we are dealing with a two-class imbalanced problem, which is challenging problem in Machine Learning [67].

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

Table 1
Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

On this account, in this section, we recall the problems that the imbalanced distribution of instances can produce (Section 4.1). Afterwards, we present the evaluation criteria that need to be considered in this framework (Section 4.2). Finally, we review some of the approaches considered to tackle this problem in the specialized literature (Section 4.3).

4.1. The class imbalance problem

A two-class classification problem is said to be imbalanced whenever the number of instances from both classes are not nearly the same, that is, one of the classes is under-represented. This fact produces a number of difficulties in the learning process and usually worsens the recognition rate of the minority (also named as positive) class [34]. In fact, the minority class is usually the most interesting one from the point of view of the learning task [29].

The problem with imbalanced data-sets comes from the fact that standard classifier learning algorithms usually fail because they are designed to maximize accuracy rate (the number of correctly classified examples). As a consequence, they are biased toward the majority (negative) class, because being easier to learn it has a greater impact on the accuracy [33]. As a result, positive instances might be treated as noise and ignored, since general rules predicting the majority class produce better accuracy rates. An imbalanced data-set does not imply an added difficulty by itself [59,29,22] (the classes can be easily separable). However, when dealing with real-world problems it usually implies the appearance of several difficulties that hinder the classifier learning. For instance, small sample size [34], overlapping [27] or small disjuncts [65], whose analysis can bring about new research directions on the topic [46]. These difficulties are amplified in high-dimensional problems [68]. Additionally, novel problems arise when dealing with class imbalance for data streams [51] in non-stationary scenarios [62].

4.2. Performance evaluation in imbalanced domains

The evaluation of the performance of a classifier is a key issue both to guide its modeling and to properly assess its quality with respect to other classifiers dealing with the same problem. When addressing a two-class problem, the results of the correctly and incorrectly classified examples of each class can be stored in a confusion matrix (Table 1).

Although historically accuracy rate has been the most commonly used measure to evaluate the performance of classifiers (Eq. (14)), it is not suitable when facing a problem with an imbalanced class distribution. This is due to the fact that accuracy rate weights the influence of the classes depending on the number of instances, and hence classes with more instances have more influence on it, which is undesirable in an imbalanced scenario.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}. \tag{14}$$

On this account, other measures need to be considered in this framework, which take the into account the performance for each class independently. From the confusion matrix (Table 1), different measures evaluating the performance over each class independently can be deduced:

- *True positive rate* (also known as *Sensitivity*) $\mathrm{TP}_{rate} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$.
- *True negative rate* $\mathrm{TN}_{rate} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}$.
- *False positive rate* $\mathrm{FP}_{rate} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$.
- *False negative rate* $\mathrm{FN}_{rate} = \frac{\mathrm{FN}}{\mathrm{TP+FN}}$.

Nonetheless, these measures on their own are still inadequate, because they only consider one of the classes. There are two commonly considered measures that can be used in this scenario:

- Geometric Mean (GM) [2]: It considers a balancing between the accuracy over the instances of the minority and majority classes at the same time (Eq. (15)) being appropriate to deal with the class imbalance problem [26,23,24].

$$\mathrm{GM} = \sqrt{\mathrm{TP}_{rate} \cdot \mathrm{TN}_{rate}}. \tag{15}$$

- Area Under the ROC Curve (AUC) [32]: Receiver Operating Characteristic (ROC) graphic [7] combines the measures obtained for each class to produce a valid evaluation criterion. ROC graphic allows one to visualize the trade-off between TP_{rate} (benefits) and FP_{rate} (costs), evidencing that increasing the number of true positives without also increasing the number of false positives is not possible for any classifier. Area Under the ROC Curve (AUC) [32] corresponds to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUC provides a scalar measure of the performance of a classifier and it has been widely used in imbalanced domains [27,26,54]. AUC measure is computed as the area of the ROC curve:

$$\mathrm{AUC} = \frac{1 + \mathrm{TP}_{rate} - \mathrm{FP}_{rate}}{2}. \tag{16}$$

4.3. Solutions for the class imbalance problem

Due to the importance of the imbalanced data-sets problem, many techniques have been developed aiming at overcoming it. These approaches can be categorized into four groups [22]:

1 Algorithm level approaches (*internal*): The methods in this category consist of adapting existing classifier learning algorithms to tackle the class imbalance by biasing the learning procedure towards the minority class [44,2]. Their drawback is that they require special knowledge of both the classifier and the application domain.
2 Data level (*external*): These methods balance the class distribution by data resampling [3,18]. Hence, they try to avoid the effects caused by class imbalance using a preprocessing step, which makes them independent of the classifier used, that is, more versatile. Well-known methods in this category are random undersampling, oversampling and Synthetic Minority Oversampling Technique) (SMOTE) [11].
3 Cost-sensitive learning: This category falls between data and algorithm level categories. It requires both data level transformations (adding costs to instances) and algorithm level modifications (by modifying the learning process to accept costs) [13,70]. The classifier is biased towards the positive class assigning higher misclassification costs to this class and trying to minimize the total cost of both classes. An important drawback of these methods is that they need to define the costs for each class, which are hardly ever available in classification data-sets.
4 Ensemble-based approaches [66]: These techniques have recently arisen as a new solution for the class imbalance problem with good results [22]. They usually combine an ensemble learning algorithm with one of the previous techniques, and more specifically, data level and cost-sensitive ones. In the case of data level approaches, a selected pre-processing algorithm is used before training each classifier of the ensemble to balance

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

the class distribution [12,56,23]. Cost-sensitive ensembles work on the basis of costs in the ensemble learning algorithm, cost-sensitive evaluation can be used locally for each base classifier [52] or globally as an evaluation metric [58].

The last category is the one in which we have focused to tackle the problem of the breast cancer malignancy classification due to the good performance shown by ensemble-based approaches, which have shown to outperform other models such as the commonly used data level methods [22,56]. More specifically, we focus on EUSBoost model [23], which combines Evolutionary Undersampling [26] with AdaBoost.M2 algorithm [21] (explained in the next section).

5. Evolutionary undersampling boosting

In this section, we explain our proposal to overcome the class imbalance problem in breast cancer malignancy grading with ensembles of classifiers. The most common approach to address the class imbalance with ensembles is the introduction of a data preprocessing step which balances the data distribution. These methods are more versatile than those based on cost-sensitive ensembles [58] because the setting of the costs is avoided. The preprocessing-based ensemble methods can be further divided into Bagging-, Boosting-, and Hybrid-based ensembles, depending on the ensemble learning algorithm in which they are based [22]. An extensive empirical analysis of ensemble solutions for class imbalance was carried out in [22], where both Boosting [21] and Bagging [8] in combination with preprocessing techniques achieved the best results. Among the methods studied in [22], RUSBoost [56] was shown to be one of the most accurate approaches (it is based on random undersampling in combination with Boosting). However, a modification of this method called EUSBoost [23] was able to outperform RUSBoost by introducing the usage of evolutionary undersampling (EUS) [26], and therefore improving the accuracy and diversity of the base classifiers. For this reason, we consider this method for our proposal and we recall its operating procedure in the following subsections.

First in Section 5.1, we present EUSBoost algorithm as an extension of Boosting using EUS with a modified fitness function. Afterwards, we recall EUS algorithm in Section 5.2. Finally, the modification of the original fitness function used in EUS in order to promote diversity is presented in Section 5.3.

5.1. Combining boosting and evolutionary undersampling

Before introducing the hybridization of Boosting and EUS, we should recall how Boosting [21] algorithm works. In Boosting classifiers are learned serially using the whole training set in all the base classifiers. However, the instances are weighted (starting with equal weights in the first iteration) and more focus is give to difficult instances after each round, aiming at correctly classifying in the current iteration those examples that were incorrectly classified in the previous one. EUSBoost is based on AdaBoost.M2 [21], which has been widely employed in imbalanced domains in combination with data level techniques. The main advantage of this AdaBoost variant is that it takes advantage of the confidences given by the base classifiers in the weight update.

The combination of EUS with Boosting algorithm is direct, yet effective. The idea of RUSBoost and other Boosting-based algorithms [22] is followed, introducing the undersampling process in each loop of AdaBoost.M2 algorithm in order to balance the class distribution. In the case of EUSBoost, as well as in the rest of the approaches, the weights of the instances are only used in the learning of the base classifiers, whereas the undersampling process (using EUS) is carried out independently of them. The complete EUSBoost algorithm is presented in Algorithm 1. Notice that the new steps with respect to AdaBoost.M2 are the 7th and 8th, while the 9th is modified.

- *Step* 7 – EUS is introduced, returning a new data-set (S') which considers all the minority class instances and the selected ones from the majority class.
- *Step* 8 – The weights for the new data-set are computed.
- *Step* 9 – The classifier is trained. Even though the original data-set is maintained, those instances not present in the undersampled data-set have no weight, being ignored in the learning of the classifier.

Algorithm 1. EUSBoost, EUS embedded in AdaBoost.M2

Input: Training set $S = \{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, N$; and $y_i \in \{c_1, c_2\}$; T: Number of iterations; I: Weak learner

1: $D_1(i) \leftarrow 1/N$ for $i = 1, \ldots, N$ {D is the weight distribution for the instances}
2: $w^1_{i,y} \leftarrow D_1(i)$ for $i = 1, \ldots, N$, $y \neq y_i$ {w, W and q_t are weights computed from D that used along the algorithm}
3: **for** $t = 1$ to T **do**
4: $W^t_i \leftarrow \sum_{y \neq y_i} w^t_{i,y}$
5: $q_t(i, y) \leftarrow \frac{w^t_{i,y}}{W^t_i}$ for $y \neq y_i$
6: $D_t(i) \leftarrow \frac{W^t_i}{\sum_{i=i}^{N} W^t_i}$
7: S' = EvolutionaryUndersampling(S);
8: $D'_t(k) \leftarrow \begin{cases} \frac{W^t_i}{\sum_{x_i \in S'} W^t_i} & \text{if } x_i \in S' \\ 0 & \text{otherwise} \end{cases}$
9: $h_t \leftarrow I(S, D'_t)$
10: $\epsilon_t \leftarrow \frac{1}{2}\sum_{i=1}^{N} D_t(i)\left(1 - h_t(\mathbf{x}_i, y_i) + \sum_{i, y \neq y_i} q_t(i, y) h_t(\mathbf{x}_i, y)\right)$
11: $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$ {β_t is the weight assigned to the t^{th} classifier}
12: $w^{t+1}_{i,y} = w^t_{i,y} \cdot \beta_t^{\frac{1}{2}(1 + h_t(\mathbf{x}_i, y_i) - h_t(\mathbf{x}_i, y))}$ for $i = 1, \ldots, N$, $y \neq y_i$
13: **end for**

Output: Boosted classifier: $H(x) = \arg\max_{y \in \mathbb{C}} \sum_{t=1}^{T} \ln\left(\frac{1}{\beta_t}\right) h_t(x, y)$, where h_t, β_t (with $h_t(x, y) \in [0, 1]$) are the classifiers and their assigned weights, respectively

The usage of EUS in the imbalance framework when constructing the ensemble allows one to better control the randomness behind the ensemble in such a way that the accuracy over the minority class can be boosted. Due to the initial randomness of the solutions of EUS, the resulting data subsets usually differ from one execution to another. EUSBoost benefits from this instability, since it helps maintaining the diversity (classifiers trained with identical data-sets are not useful to construct ensembles). However, a key factor of EUSBoost is the definition of a new fitness function for EUS, which takes into account the diversity of the instance subset obtained with respect to the already used ones in such a way that the final diversity of the ensemble is improved. This procedure is explained in Section 5.3, after recalling EUS in Section 5.2.

5.2. Evolutionary undersampling

EUS [26] is an evolutionary prototype selection algorithm adapted to work in imbalanced domains (it uses an appropriate fitness function). Prototype selection [25] is a sampling process aiming at reducing the reference set for the nearest neighbor (1NN) classifier in order to improve its accuracy and reduce the storage necessity. However, in an imbalanced scenario the objective differs, since the balance of the data distribution gains importance.

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

On this account, EUS tries to obtain a useful undersampled data-set whose search is guided by a genetic algorithm. Initially, several randomly undersampled data subsets are created, which are then evolved until the currently best undersampled data-set cannot be further improved in terms of the fitness function. This algorithm has already shown its usefulness in real-world applications [16].

Likewise in every evolutionary method, the way in which the solutions are represented by means of chromosomes is an important issue. In EUS, a binary vector is used to represent each solution, where each gene (binary value) represents the presence or absence of the corresponding instance in the data-set. Although all of instances could be codified in the chromosome, the search space can be reduced by only considering the majority class instances, whereas all the minority class instances are always introduced in the undersampled data-set. Therefore, a chromosome is represented as follows:

$$V = (v_{x_1}, v_{x_2}, v_{x_3}, v_{x_4}, \ldots, v_{x_{n^-}}), \tag{17}$$

where v_{x_i} takes the values 0 or 1, indicating whether instance x_i is included or not in the data-set (n^- stands for the number of majority class instances).

In the evolutionary process, chromosomes are ranked using a fitness function, which in the case of EUS takes into account the balancing between both classes and the expected performance with the selected data subset [26]. In order to estimate this performance a hold-one-out technique is used with 1NN classifier and it is measured by the GM. Finally, the fitness function of EUS is as follows:

$$\text{fitness}_{\text{EUS}} = \begin{cases} \text{GM} - \left|1 - \dfrac{n^+}{N^-} \cdot P\right| & \text{if } N^- > 0 \\ \text{GM} - P & \text{if } N^- = 0, \end{cases} \tag{18}$$

where n^+ is the number of minority class instances, N^- is the number of majority class instances selected and P is the penalization factor accounting for the importance given to the balance between both classes (whose recommended value is 0.2).

In order to perform the search, the well-known CHC algorithm [17] is used due to its good balance between exploration and exploitation. CHC is an elitist genetic algorithm using the heterogeneous uniform cross-over (HUX) to combine two chromosomes (exactly half of the different genes are interchanged). An incest prevention mechanism is also considered where two parents are only recombined if their Hamming distance is greater than the threshold (initially $L/4$, being L the length of the chromosome); the threshold is reduced by one when no parents are recombined. In this genetic algorithm no mutation is applied, but when the recombined chromosomes are not able to improve their parents and the threshold reaches zero, the whole population (except for the best chromosome) are reinitialized. Reinitialization consists of using the best chromosome as a template, randomly changing 35% of its genes.

In the case of EUS, the original HUX is modified to decrease the probability of including instances in the data-set in such a way that a good reduction rate is reached. To do so, each time HUX switches a gene on, it is switched off with a certain probability (the recommended value is 0.25).

5.3. Promoting diversity: adjusting the fitness function of evolutionary undersampling

In order to improve the diversity of the base classifiers in EUSBoost, the original fitness function of EUS is modified. Only considering the randomness in EUS as a source of diversity may not be enough to provide different subsets of instances in each iteration of Boosting. Hence, diversity among data-sets is promoted in a supervised manner, which is not usually done.

Diversity in classifier ensembles refers to the fact that they should be composed of base classifiers giving different outputs so that their combination can lead to significant improvements. Diversity is a key factor in ensembles and has been widely studied in the literature [10,64] even though no direct relation has been found between diversity and accuracy. However, in the imbalanced scenario diversity gains importance. In [64] it was found that there exists a relation between diversity and single-class performance measures, having a positive impact on the minority class classification, but also on global performance measures such as AUC.

Since EUS is used as a preprocessing algorithm, the diversity of the outputs cannot be directly promoted. For this reason, the diversity among solutions is considered, that is, chromosomes that are different from the best chromosomes in previous iterations are preferred. Hence, it is assumed that base classifiers learned from data-sets with more different instances are more diverse (which is also assumed in Bagging, but in this case it is forced). To do so, a new fitness function is introduced, modifying the original evaluation procedure of EUS (Eq. (18)).

First, diversity between solutions should be measured, which is done using the Q-statistic [69], which has been widely applied in classifier ensembles [42,64]. Recall that this measure is applied to compute the diversity between two solutions (Eq. (17)). Having two binary vectors (V_i, V_j), the Q-statistic is computed as follows:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{19}$$

where N^{ab} stands for number of instances with value a in the first vector and with value b in the second (if $a = b$ both data-sets agree including or not the instance). The value of the statistic ranges from -1 to 1. Lower values of Q indicates greater diversity ($Q_{i,j} = 0$ means that both vectors are statistically independent).

The Q-statistic is a pairwise measure [41], but the diversity between a candidate chromosome and the previously used best ones are computed. In order to aggregate all the pairwise values the maximum of all $Q_{i,j}$ is considered. In this way, the candidate instance subset that is the most dissimilar with respect to all the previous data-sets is considered. Therefore, being V_j the candidate solution to be evaluated, and V_i, $i = 1, \ldots, t$ (recall that t is the current iteration) all the previously used solutions, we compute the global diversity Q as:

$$Q = \max_{i=1,\ldots,t} Q_{i,j}. \tag{20}$$

Once the evaluation of the diversity has been defined, EUS's fitness function is modified as follows:

$$\text{fitness}_{\text{EUS}_Q} = \text{fitness}_{\text{EUS}} \cdot \frac{1.0}{\beta} \cdot \frac{10.0}{\text{IR}} - Q \cdot \beta, \tag{21}$$

where $\text{fitness}_{\text{EUS}}$ is the original fitness function (Eq. (17)), IR is the imbalance ratio (the number of negative class examples divided by the number of positive class examples), Q is the global Q-statistic and β is a weight factor changing in each iteration:

$$\beta = \frac{N - t - 1}{N}. \tag{22}$$

Notice that in Eq. (21) the Q term is subtracted to maximize the diversity. Furthermore, in the first iteration of the ensemble ($t = 1$), EUS is executed in its original form (using Eq. (18)), since there are no vectors to compare with the actual candidate solution (for more explanation on this fitness function we refer the reader to the original paper of EUSBoost [23]).

6. Experimental analysis

In this section we develop an exhaustive experimental study to check the usefulness of EUSBoost ensemble for the classification of

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

Table 2
Features extracted for malignancy classification.

100× magnification features	– Area of the groups – Number of groups – Dispersion	
400× Magnification features	Binary features	– Area of the nucleus, – Nucleus perimeter, – Convexity, eccentricity, – Coordinates of the nucleus centroid – Nucleus orientation, – Horizontal and vertical projections
	Momentum features	– $\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6, \varphi_7$
	Histogram features	– Mean, standard deviation, – Skew, energy, entropy
	Textural features	– Inverse difference, correlation – Average gray level
	Color histogram Features	– Mean, standard deviation, – Skew, energy, width

breast cancer malignancy grading. Moreover, we compare the three image segmentation techniques described in Section 3 and discuss which one is the most appropriate to deal with this problem. The experimental analysis has two main research objectives:

- To compare EUSBoost to a number of state-of-the-art methods dedicated to the imbalanced classification problem and check whether it can be useful in the process of designing a clinical decision support system for real-life application.
- To investigate whether there is a difference in the discriminative power of the features obtained through the three different feature extraction methods considered, and to check which one performs better with the proposed ensemble classifier.

In the following subsections, we will describe the used dataset, bring details about the selected classification methods included in the comparison and their parameters, discuss the set-up of the experiments and present the obtained results together with the corresponding discussion.

6.1. Dataset

In this study we have used a database of FNA slides. The data was collected at the Department of Pathology and Oncological Cytology, Medical University of Wrocław, Poland. Currently the dataset consists of 341 images with a resolution of 96 dots per inch (dpi) and a size of 764 × 572 pixels. All of the images were taken with Olympus BX 50 microscope with mounted CCD-IRIS camera connected to a PC computer with MultiScan Base 08.98 software. Prior digitalization, the FNA slides were stained with the Haematoxylin and Eosin technique (HE) which yielded purple and black stain for nuclei, shades of pink for cytoplasm and orange/red for red blood cells.

The organization of the images in the dataset follows the requirements of the diagnostic process in which two types of slides are taken into consideration. The first type (167 images) are the slides recorded in low magnification (100×) and the second type (174 images) are the slides recorded in high magnification (400×). The low magnification images are used to define features related to the degree of structural differentiation (see Section 2), that is, cells' ability to form groups. 400× magnification images are the base for the calculation of features that reflect cells' polimorphy and mitotic count. An extensive description of these features can be found in Krawczyk et al. [40].

During examination there might be more than one high magnification image taken from one slide. This is the reason why in the database we sometimes have more than one 400× image for a 100× image. From the diagnostic point of view, this is caused by the fact that there are more than one suspicious region in the 100× image. Here, for the purpose of this study we have separated this cases and treated them as different cancer occurrences. For the classification purposes we have used a pair of images that consisted of one 100× image and one 400× image and based on the features computed (for feature extraction, please refer to Section 6.2), the malignancy grade was automatically assigned.

In the dataset there are occurrences of intermediate (G2) and high (G3) malignancy grades. There are no images of low malignancy because there were no such cases at the Medical University of Wrocław, Poland since 2004. Here, we collected 268 images belonging to the G2 class and 73 to the G3 class which means that there are 137 cases of intermediate malignancy and 39 cases of high malignancy. This imbalanced dataset makes the classification scheme more difficult and was the motivation to perform a series of studies to achieve the best classification possible using ensembles designed to tackle this problem.

6.2. Features

All the images in the database were segmented according to the description in Section 3 and based on these segmentations a set of features was calculated. In this paper, we extracted 32 features that consisted of three features calculated from the low magnification images (100× magnification features) and 29 features from high magnification images (400× magnification features). All the features were determined according to the description of Krawczyk et al. [40] and they are summarized in Table 2.

For images recorded with 100× magnification we have calculated three features that represent cells' tendency to form groups. These are very important features during cancer diagnosis and therefore are taken into consideration in this study. For 400× images we have calculated 29 features in 5 categories. First category corresponds to binary features that are calculated based on the segmented binary image. These features allow us to estimate the polymorphy of the nuclei. The next category is formed of RST-invariant features (φ_1–φ_7) that are calculated based on the normalized central moments. Other two groups of features are histogram based ones and color histogram features, which are based on image histogram. The last group encompasses textural features, which are computed from a gray-level co-occurrence matrix and represent changes in nuclei texture. Textural information is important as it describes a chromatin changes inside each nucleus.

6.3. Set-up

In order to put the results obtained into context, we compare our method with several state-of-the-art algorithms dedicated to binary and imbalanced classification [22]. The list of used learning algorithms is given in Table 3, while their parameters are given

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

Table 3
Short description of classification models used in experiments.

Abbr.	Model	Short description
Standard ensembles		
BGG	Bagging	Standard Bagging, resampling with replacement, bag size equal to original data size
BST	AdaBoost.M2	Boosting using confidence estimates
Bagging-based ensembles for imbalanced problems		
UNB	UnderBagging [2]	Bagging with undersampling of the majority class, doubles the number of positive examples
RBB	Roughly Balanced Bagging [30]	Bagging with undersampling of the majority class, number of negative examples varies in each bag
OVB	OverBagging [63]	Bagging with oversampling of the minority class, doubles the number of negative examples
SBG	SMOTEBagging [63]	Bagging with varied SMOTE quantity in each bag
IIV	IIVotes [5]	IVotes [9] with SPIDER (strong) oversampling in each iteration
Boosting-based ensembles for imbalanced problems		
ADC	AdaC2 [58]	AdaBoost with cost outside exponent
RUB	RUSBoost [56]	AdaBoost.M2 with random undersampling in each iteration
SBO	SMOTEBoost [12]	AdaBoost.M2 with SMOTE in each iteration
MBO	MSMOTEBoost [31]	AdaBoost.M2 with MSMOTE in each iteration
Hybrid ensembles for imbalanced problems		
EAE	EasyEnsemble [45]	UnderBagging learned with AdaBoost in each bag
BAL	BalanceCascade [45]	EasyEnsemble with removing majority examples in each iteration
Proposed method		
EUB	EUSBoost	AdaBoost.M2 with evolutionary undersampling

in Table 4. Each ensemble uses C4.5 classifier as base learner. The parameter values were established through a grid-search procedure so that their performance is expected to be the best possible one.

All experiments were done with the usage of combined 5×2 cv F test [1], which allows one to simultaneously perform cross-validation and to assess the statistical significance of the obtained results. Statistical significance level $\alpha = 0.05$ is assumed.

6.4. Experimental results

Results of the experiments, with the respect to sensitivity, GM and AUC for considered classifiers and segmentation methods are presented in Table 5. Detailed results of the pairwise combined 5×2 cv F test for each segmentation method and used performance metrics are given in Tables 6–8.

Table 4
Parameters of used pre-processing and classification methods. The parameter values were established through a grid-search procedure.

Algorithm	Parameters
C4.5	Prune = true; confidence level = 0.25; Minimum number of item-sets-per-leaf = 2; Confidence = Laplace smoothing
SMOTE/MSMOTE	No. of neighbors = 5; quantity = balance; distance = Heterogeneous value difference metric
SPIDER	No. of neighbors = 5
BGG	No. of bags = 16
BST	No. of iterations = 10
UNB	No. of bags = 16
RBB	No. of bags = 12
OVB	No. of bags = 20
SBG	No. of bags = 12
ADC	$C_{min} = 1$, $C_{maj} = 1/IR$
RUB	No. of iterations = 10
SBO	No. of iterations = 12
MBO	No. of iterations = 12
EAE	No. of bags = 4
BAL	No. of iterations = 10
EUB	Population size = 30; number of evaluations = 5000; Probability of inclusion HUX = 0.25; evaluation measure = GM; Selection type = Majority; distance function = Euclidean; Balancing = true; $P = 0.2$; no. of iterations = 10

6.5. Discussion

The experimental analysis allowed us to thoroughly examine the usefulness of the used ensemble classifiers for imbalanced classification in real clinical decision support system. Additionally, we were able to analyze the correlation between the used segmentation method and classification quality. Below, we present the main conclusions that can be drawn from this experiment.

Firstly, looking at the performance of the different ensemble methods, standard approaches, like bagging or boosting, fail to deliver a satisfactory performance in this task, regardless of the feature set or measure used. Therefore, it can be concluded that the difficulty is not only connected with the uneven distribution, but also with presence of difficulties embedded in the nature of data, which cannot be directly addressed with classical models.

Another trend that can be observed is connected with the used type of ensemble classifiers. We can look at the applied methods from two perspectives: based on their learning algorithm (here we can distinguish bagging-based, boosting-based or hybrid approaches) or on their methodology for handling the class imbalance (undersampling, oversampling and cost-sensitive methods).

When grouping the methods according to their learning paradigm, one may clearly observe that hybrid-based approaches (EasyEnsemble and BalanceCascade) return the worst results for all of measures used. Previous literature reports about their performance on benchmark datasets showed, that they are competitive to other classifiers. however, in this real-life medical problem they are far behind their rivals. This may indicate, that the analyzed data has some difficult samples (rare or overlapping), that cannot be efficiently tackled by pure classifier-level approaches. This is further backed-up by the observation that methods which use pre-processing algorithms achieve much better results. Bagging-based and boosting-based algorithms in general obtain similar results, which confirms the findings reported in a thorough review on these models [22]. The main differences are related to the pre-processing technique embedded in the ensemble, not the ensemble forming procedure itself. However, one should note that the most effective method is boosting-based.

Regarding the analysis of the results looking at the way ensembles handle the class imbalance problem, the worst performance is

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

Table 5
Experimental results (sensitivity, GM and AUC), with the respect to used image segmentation and classification methods.

Segmentation	Measure	BGG	BST	UNB	RBB	OVB	SBG	IIV	ADC	RUB	SBO	MBO	EAE	BAL	EUB
FCM	SEN	72.87	72.92	88.32	90.18	85.04	86.10	88.15	88.22	89.73	87.39	88.21	85.29	86.08	**92.19**
	GM	76.28	78.82	90.13	91.51	88.48	88.82	89.03	91.07	91.74	89.03	88.79	89.38	87.68	**92.28**
	AUC	78.86	80.23	90.48	92.30	88.98	90.02	90.78	91.26	92.38	89.39	88.96	90.52	88.26	**93.88**
LS	SEN	75.98	75.11	90.28	92.85	88.09	88.43	90.97	91.04	92.12	90.07	90.52	89.00	88.23	**94.53**
	GM	80.31	82.04	92.74	94.37	90.76	91.26	91.87	93.05	94.28	91.87	92.00	91.22	90.84	**95.73**
	AUC	81.02	83.17	92.89	94.83	90.98	92.31	92.31	93.36	95.09	91.92	93.02	92.38	91.07	**96.38**
GLQ	SEN	70.47	71.63	85.30	**90.27**	83.92	85.11	86.89	88.44	**90.27**	85.92	87.03	84.11	84.68	**90.27**
	GM	71.62	73.01	87.26	**90.89**	85.36	86.38	87.12	88.68	**90.89**	86.39	87.94	85.38	86.38	**90.89**
	AUC	72.14	73.39	87.97	**90.74**	86.02	86.92	88.02	88.98	**90.74**	86.98	88.40	85.69	86.92	**90.74**

Table 6
Combined 5 × 2 cv *F* test for comparison between the proposed EUSBoost and reference methods for fuzzy c-means color segmentation. Symbol '+' stands for situation in which EUSBoost is superior, '-' if reference method is superior and '=' for classifiers without significant differences.

Hypothesis	*p*-value (SEN)	*p*-value (GM)	*p*-value (AUC)
EUB vs. BGG	+ (0.0046)	+ (0.0042)	+ (0.0068)
EUB vs. BST	+ (0.0033)	+ (0.0037)	+ (0.0054)
EUB vs. UNB	+ (0.0207)	+ (0.0184)	+ (0.0192)
EUB vs. RBB	+ (0.0375)	+ (0.0401)	+ (0.0397)
EUB vs. OVB	+ (0.0116)	+ (0.0132)	+ (0.0149)
EUB vs. SBG	+ (0.0136)	+ (0.0161)	+ (0.0169)
EUB vs. IIV	+ (0.0218)	+ (0.0299)	+ (0.0265)
EUB vs. ADC	+ (0.0257)	+ (0.0328)	+ (0.0303)
EUB vs. RUB	+ (0.0342)	+ (0.0388)	+ (0.0395)
EUB vs. SBO	+ (0.0186)	+ (0.0197)	+ (0.0188)
EUB vs. MBO	+ (0.0190)	+ (0.0210)	+ (0.0204)
EUB vs. EAE	+ (0.0128)	+ (0.0136)	+ (0.0130)
EUB vs. BAL	+ (0.0145)	+ (0.0153)	+ (0.0151)

obtained by ensembles based on over-sampling. Regardless of the approach used (simple: OverBagging; or more complex: introducing new artificial examples with SMOTE/MSMOTE/SPIDER) these methods cannot efficiently improve the recognition rate of the minority class. And what is even more interesting, these methods seem to boost the classification of the majority class (which can be deducted from a larger increase on GM with the respect to other ensembles), but do not return a satisfactory sensitivity.

On the other hand, undersampling-based methods tend to return very good performances, even when a simple balanced reduction of the majority class is carried out. Especially remarkable results are returned by Roughly Balanced Bagging and RUBoost. Roughly Balanced Bagging takes an advantage of a specific negative binomial distribution for drawing samples into bags, which introduces additional diversity and puts different emphasis on examples from the considered classes. This is especially important in the case of complex distributions, where uniformly formed bags may not capture well enough the characteristics of the considered problem. Otherwise, RUBoost uses a simple random undersampling, which makes the distribution of samples in each iteration equal to 50:50, but takes a full advantage of AdaBoost.M2 algorithm to form the ensemble. However, due to the simple method for reducing the class imbalance, it can overlook some important underlying structures in both classes and lose such information in the process.

Anyway, looking at the results there is a method which outstands with respect to the previous ones when addressing the breast cancer malignancy grading classification. The proposed EUSBoost stands superior to all other methods, as proven with all three measures used and the statistical analysis. This can be explained by the fact that it significantly extends RUBoost, which is an effective method for this problem on its own. EUSBoost also applies AdaBoost.M2 algorithm to create the ensemble, but further back it up with an guided evolutionary undersampling. This method at the same time searches for the most descriptive examples of the majority class, while maintaining a diversity thanks to the measurement of diversity using *Q*-statistic embedded in the evolutionary procedure. EUSBoost therefore can handle the imbalanced breast cancer malignancy data more efficiently by creating individually accurate and mutually complementary base classifiers. It is worth to notice that EUSBoost was originally created for datasets with high imbalance ratio. However, in this paper we have proven its usefulness for scenarios, where the difficult lies not in the big disproportion in the number of objects, but in a difficulties embedded in the data. EUSBoost is able to select the most useful samples for

Table 7
Combined 5 × 2 cv *F* test for comparison between the proposed EUSBoost and reference methods for level set active contours segmentation. Symbol '+' stands for situation in which EUSBoost is superior, '-' if reference method is superior and '=' for classifiers without significant differences.

Hypothesis	*p*-value (SEN)	*p*-value (GM)	*p*-value (AUC)
EUB vs. BGG	+ (0.0057)	+ (0.0063)	+ (0.0071)
EUB vs. BST	+ (0.0073)	+ (0.0079)	+ (0.0082)
EUB vs. UNB	+ (0.0217)	+ (0.0196)	+ (0.0208)
EUB vs. RBB	+ (0.0412)	+ (0.0423)	+ (0.0416)
EUB vs. OVB	+ (0.0141)	+ (0.0167)	+ (0.0159)
EUB vs. SBG	+ (0.0159)	+ (0.0174)	+ (0.0182)
EUB vs. IIV	+ (0.0238)	+ (0.0320)	+ (0.0291)
EUB vs. ADC	+ (0.0263)	+ (0.0342)	+ (0.0317)
EUB vs. RUB	+ (0.0403)	+ (0.0415)	+ (0.0413)
EUB vs. SBO	+ (0.0202)	+ (0.0218)	+ (0.0209)
EUB vs. MBO	+ (0.0216)	+ (0.0233)	+ (0.0221)
EUB vs. EAE	+ (0.0160)	+ (0.0165)	+ (0.0168)
EUB vs. BAL	+ (0.0151)	+ (0.0156)	+ (0.0154)

Table 8
Combined 5 × 2 cv *F* test for comparison between the proposed EUSBoost and reference methods for grey-level quantization segmentation. Symbol '+' stands for situation in which EUSBoost is superior, '-' if reference method is superior and '=' for classifiers without significant differences.

hypothesis	*p*-value (SEN)	*p*-value (GM)	*p*-value (AUC)
EUB vs. BGG	+ (0.0071)	+ (0.0083)	+ (0.0080)
EUB vs. BST	+ (0.0094)	+ (0.0106)	+ (0.0103)
EUB vs. UNB	+ (0.0228)	+ (0.0232)	+ (0.0229)
EUB vs. RBB	= (0.3482)	= (0.3561)	= (0.3507)
EUB vs. OVB	+ (0.0177)	+ (0.0186)	+ (0.0183)
EUB vs. SBG	+ (0.0138)	+ (0.0144)	+ (0.0157)
EUB vs. IIV	+ (0.0236)	+ (0.0253)	+ (0.0277)
EUB vs. ADC	+ (0.0227)	+ (0.0251)	+ (0.0248)
EUB vs. RUB	= (0.3482)	= (0.3561)	= (0.3507)
EUB vs. SBO	+ (0.0171)	+ (0.0192)	+ (0.0195)
EUB vs. MBO	+ (0.0188)	+ (0.0204)	+ (0.0197)
EUB vs. EAE	+ (0.0157)	+ (0.0182)	+ (0.0185)
EUB vs. BAL	+ (0.0228)	+ (0.0232)	+ (0.0229)

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060

creating classifiers, alleviating the drawbacks of random selection. Of course EUSBoost is more computationally expensive than RUBoost or Roughly Balanced Bagging. But in this context, we are designing a clinical decision support system for application in hospitals. Therefore, we are not as much concerned with the training time, as with the classification time and final accuracy. After training procedure EUSBoost has a response time identical to other ensemble techniques, while returning a statistically superior recognition quality.

Finally, we should analyze the results of the three used segmentation methods: fuzzy c-means color segmentation, level set active contours technique and grey-level quantization method that uses a texture information for segmentation. There is no direct correlation between the used type of classifier and segmentation methods, as the ranks of the classifiers are similar for every three sets of features. With each set of features Roughly Balanced Bagging, RUBoost and EUSBoost return the best results. However, the main difference lies in the discriminative power of the features extracted with these methods.

The best results are given by level set active contours technique, allowing EUSBoost to reach 95.43% of sensitivity, GM equal to 95.73% and AUC equal to 96.38. Ensembles based on features extracted through fuzzy c–means segmentation are on average 2%–4% inferior, regardless of the measure considered. Finally, grey-level quantization method outputs the features with the lowest usefulness for classification purposes. This is the only case, in which EUSBoost is not able to outperform Roughly Balanced Bagging and RUBoost, as all three of them display an identical performance (which is inferior to that obtained with the other two methods).

Summarizing, it can be concluded that EUSBoost trained on features extracted through level set active contours technique allows one to create highly efficient clinical decision support system for breast cancer malignancy diagnosis.

7. Conclusions

In this paper, we proposed a complete, fully automatic and highly accurate clinical decision support system based on ensemble classification. We have dealt with an imbalanced problem in which the minority class corresponded to the highest (and thus most important to detect) malignancy grade. We discussed three different methods for FNA slides segmentation and feature extraction. On their basis, we trained a novel ensemble classifier called EUSBoost. It combined a boosting scheme with evolutionary undersampling. This allowed us to perform a guided undersampling of the majority class, selecting the most important objects for the classifier training step. Additionally, by incorporating a diversity measure in the evolutionary algorithm we were able to assure that the classifiers are mutually complementary. Our method obtained the best results when features extracted from level set active contours technique were used, returning 95.43% of sensitivity, GM equal to 95.73% and AUC equal to 96.38. These excellent results allowed EUSBoost to outperform 13 state-of-the-art ensemble classifiers, which was backed-up with a thorough statistical analysis of the results.

The proposed clinical decision support system can be easily implemented on a standard computer and may be a valuable aid to the everyday physician's routine.

Acknowledgements

Bartosz Krawczyk was partially supported by the Polish National Science Center under the grant no. DEC-2013/09/B/ST6/02264.

Mikel Galar was partially supported by the Spanish Ministry of Education and Science under Project TIN2013-40765-P.

Francisco Herrera was partially supported by the Spanish Ministry of Education and Science under Project TIN2011-28488 and the Andalusian Research Plan P10-TIC-6858, P11-TIC-7765.

References

[1] E. Alpaydin, Combined 5 × 2 cv *F* test for comparing supervised classification learning algorithms, Neural Comput. 11 (8) (1999) 1885–1892.
[2] R. Barandela, R.M. Valdovinos, J.S. Sánchez, New applications of ensembles of classifiers, Pattern Anal. Appl. 6 (3) (2003) 245–256.
[3] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explor. Newsl. 6 (2004) 20–29.
[4] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[5] J. Blaszczynski, M. Deckert, J. Stefanowski, S. Wilk, Integrating selective preprocessing of imbalanced data with Ivotes ensemble, in: Volume 6086 LNAI of Lecture Notes in Computer Science, 2010, pp. 148–157.
[6] H.J.G. Bloom, W.W. Richardson, Histological grading and prognosis in breast cancer, Br. J. Cancer 11 (1957) 359–377.
[7] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[9] L. Breiman, Pasting small votes for classification in large databases and on-line, Mach. Learn. 36 (1–2) (1999) 85–103.
[10] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Inf. Fus. 6 (1) (2005) 5–20, Diversity in Multiple Classifier Systems.
[11] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[12] N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer, Smoteboost: improving prediction of the minority class in boosting, in: Lecture Notes in Artificial Intelligence, vol. 2838, 2003, pp. 107–119.
[13] N.V. Chawla, D.A. Cieslak, L.O. Hall, A. Joshi, Automatically countering imbalance and its empirical relationship to cost, Data Min. Knowl. Discov. 17 (2008) 225–252.
[14] H.D. Cheng, X. Cai, X. Chen, L. Hu, X. Lou, Computer-aided detection and classification of microcalcifications in mammograms: a survey, Pattern Recognit. 36 (12) (2003) 2967–2991.
[15] B. Cyganek, One-class support vector ensembles for image segmentation and classification, J. Math. Imaging Vis. 42 (2–3) (2012) 103–117.
[16] D.J. Drown, T.M. Khoshgoftaar, N. Seliya, Evolutionary sampling and software quality modeling of high-assurance systems, IEEE Trans. Syst. Man Cybern. A: Syst. Hum. 39 (5) (2009) 1097–1107.
[17] L.J. Eshelman, The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination, in: J.E. Gregory, Rawlins (Eds.), Foundations of Genetic Algorithms, Morgan Kaufmann, San Francisco, CA, 1991, pp. 265–283.
[18] A. Fernández, S. García, M.J. del Jesus, F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets Syst. 159 (18) (2008) 2378–2398.
[19] P. Filipczuk, T. Fevens, A. Krzyżak, R. Monczak, Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies, IEEE Trans. Med. Imaging 32 (12) (2013) 2169–2178.
[20] P. Filipczuk, B. Krawczyk, M. Woźniak, Classifier ensemble for an effective cytological image analysis, Pattern Recognit. Lett. 34 (14) (2013) 1748–1757.
[21] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.
[22] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Systems Man Cybern. C: Appl. Rev. 42 (4) (2012) 463–484.
[23] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern Recognit. 46 (12) (2013) 3460–3471.
[24] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, Empowering difficult classes with a similarity-based aggregation in multi-class classification problems, Inf. Sci. 264 (0) (2014) 135–157.
[25] S. García, J. Derrac, J. Cano, F. Herrera, Prototype selection for nearest neighbor classification: taxonomy and empirical study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 417–435.
[26] S. García, F. Herrera, Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy, Evol. Comput. 17 (2009) 275–306.
[27] V. García, R. Mollineda, J. Sánchez, On the *k*-NN performance in a challenging scenario of imbalance and overlapping, Pattern Anal. Appl. 11 (2008) 269–280.
[28] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621.
[29] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[30] S. Hido, H. Kashima, Y. Takahashi, Roughly balanced bagging for imbalanced data, Stat. Anal. Data Min. 2 (5-6) (2009) 412–426.
[31] S. Hu, Y. Liang, L. Ma, Y. He, Msmote: improving classification performance when training data is imbalanced., in: Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on, vol. 2, 2009, pp. 13–17.

[32] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (3) (2005) 299–310.
[33] K. Huang, R. Zhang, X.-C. Yin, Learning imbalanced classifiers locally and globally with one-side probability machine, Neural Process. Lett. 41 (3) (2015) 311–323.
[34] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. 6 (2002) 429–449.
[35] Ł. Jeleń, Computerized Cancer Malignancy Garding of Fine Needle Aspirates, Concordia University, 2009 (PhD thesis).
[36] Ł. Jeleń, T. Fevens, A. Krzyżak, Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies, Appl. Math. Comput. Sci. 18 (1) (2008) 75–83.
[37] G.J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, New Jersey, 1995.
[38] M. Kowal, P. Filipczuk, Nuclei segmentation for computer-aided diagnosis of breast cancer, Appl. Math. Comput. Sci. 24 (1) (2014) 19–31.
[39] B. Krawczyk, P. Filipczuk, Cytological image analysis with firefly nuclei detection and hybrid one-class classification decomposition, Eng. Appl. Artif. Intell. 31 (2014) 126–135.
[40] B. Krawczyk, Ł. Jeleń, A. Krzyżak, T. Fevens, One-class classification decomposition for imbalanced classification of breast cancer malignancy data, in: ICAISC (1), 2014, pp. 539–550.
[41] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2003) 181–207.
[42] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Anal. Appl. 6 (2003) 22–31.
[43] C. Li, C. Xu, C. Gui, M.D. Fox, Level set evolution without re-initialization: a new variational formulation, in: IEEE Proceedings of Conference on Computer Vision and Pattern Recognition 2005, 2005, pp. 430–436.
[44] Y. Lin, Y. Lee, G. Wahba, Support vector machines for classification in nonstandard situations, Mach. Learn. 46 (2002) 191–202.
[45] X. Liu, J. Wu, Z. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. B: Cybern. 39 (2) (2009) 539–550.
[46] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[47] J. Malek, A. Sebri, S. Mabrouk, K. Torki, R. Tourki, Automated breast cancer diagnosis based on GVF-snake segmentation, wavelet features extraction and fuzzy classification, J. Signal Process. Syst. 55 (1-3) (2009) 49–66.
[48] M. Moghbel, M. Mashohor, A review of computer assisted detection/diagnosis (CAD) in breast thermography for breast cancer detection, Artif. Intell. Rev. 39 (4) (2013) 305–313.
[49] M.B. Nagarajan, M.B. Huber, T. Schlossbauer, G. Leinsinger, A. Krol, Classification of small lesions on dynamic breast MRI: integrating dimension reduction and out-of-sample extension into {CADx} methodology, Artif. Intell. Med. 60 (1) (2014) 65–77.
[50] S. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations, J. Comput. Phys. 79 (1988) 12–49.
[51] S. Pan, J. Wu, X. Zhu, C. Zhang, Graph ensemble boosting for imbalanced noisy graph stream classification, IEEE Trans. Cybern. 45 (5) (2015) 940–954.
[52] W. Penar, M. Woźniak, Cost-sensitive methods of constructing hierarchical classifiers, Expert Syst. 27 (3) (2010) 146–155.
[53] T.W. Ridler, S. Calvard, Picture thresholding using an iterative selection, IEEE Trans. System Man Cybern. 8 (1978) 630–632.
[54] J.A. Sáez, J. Luengo, F. Herrera, Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification, Pattern Recognit. 46 (1) (2013) 355–364.
[55] R.W. Scarff, H. Torloni, Histological typing of breast tumors. international histological classification of tumours, World Health Organ. 2 (2) (1968) 13–20.
[56] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Rusboost: a hybrid approach to alleviating class imbalance, IEEE Trans. Systems Man Cybern. A: Syst. Hum. 40 (1) (2010) 185–197.
[57] J.A. Sethian, D. Adalsteinsson, An overview of level set methods for etching, deposition, and lithography development, IEEE Trans. Semicond. Manuf. 10 (1) (1997) 167–184.
[58] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognit. 40 (12) (2007) 3358–3378.
[59] Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recognit. Artif. Intell. 23 (4) (2009) 687–719.
[60] N. Theera-Umpon, Patch-based white blood cell nucleus segmentation using fuzzy clustering, ECTI Trans. Electr. Eng. Electron. Commun. 3 (1) (2005) 15–19.
[61] M. Velikova, P.J.F. Lucas, M. Samulski, N. Karssemeijer, On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks, Artif. Intell. Med. 57 (1) (2013) 73–86.
[62] S. Wang, L.L. Minku, X. Yao, Resampling-based ensemble methods for online class imbalance learning, IEEE Trans. Knowl. Data Eng. 27 (5) (2015) 1356–1368.
[63] S. Wang, X. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 – Proceedings, 2009, pp. 324–331.
[64] S. Wang, X. Yao, Relationships between diversity of classification ensembles and single-class performance measures, IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 206–219.
[65] G.M. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, J. Artif. Intell. Res. 19 (2003) 315–354.
[66] M. Woźniak, M. Graña, E. Corchado, A survey of multiple classifier systems as hybrid systems, Inf. Fus. 16 (2014) 3–17.
[67] Q. Yang, X. Wu, 10 challenging problems in data mining research, Int. J. Inf. Technol. Decis. Mak. 5 (4) (2006) 597–604.
[68] H. Yu, J. Ni, An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (4) (2014) 657–666.
[69] G. Yule, On the association of attributes in statistics, Philos. Trans. A 194 (1900) 257–319.
[70] S. Zhang, L. Liu, X. Zhu, C. Zhang, A strategy for attributes selection in cost-sensitive decision trees induction, in: IEEE 8th Int. Conf. on Computer and Information Technology Workshops, 2008, pp. 8–13.

Please cite this article in press as: B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.08.060