Tackling the Problem of Data Imbalancing for Melanoma Classification

First Author Name[1], Second Author Name[1] and Third Author Name[2]
[1]*Institute of Problem Solving, XYZ University, My Street, MyTown, MyCountry*
[2]*Department of Computing, Main University, MySecondTown, MyCountry*
{f_author, s_author}@ips.xyz.edu, t_author@dc.mu.edu

Keywords: IMBALANCED, CLASSIFICATION, MELANOMA, DERMOSCOPY

Abstract: Malignant melanoma is the most dangerous type of skin cancer, yet melanoma is the most treatable kind of cancer when diagnosed at an early stage. In this regard, Computer-Aided Diagnosis systems based on machine learning have been developed to discern melanoma lesions from benign and dysplastic nevi in dermoscopic images. Similar to a large range of real world applications encountered in machine learning, melanoma classification faces the challenge of imbalanced data, where the percentage of melanoma cases in comparison with benign and dysplatic cases is far less. This article analyzes the impact of data balancing strategies at the training step. Subsequently, Over-Sampling (OS) and Under-Sampling (US) are extensively compared in both feature and data space, revealing that NearMiss-2 (NM2) outperform other methods achieving Sensitivity (SE) and Specificity (SP) of 91.2% and 81.7%, respectively. More generally, the reported results highlight that methods based on US or combination of OS and US in feature space outperform the others.

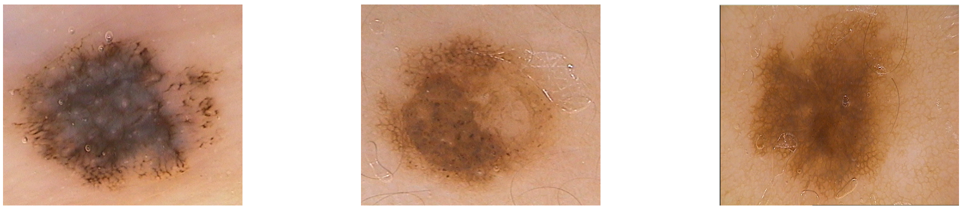
1 INTRODUCTION

Malignant melanoma is the deadliest type of skin cancer, accounting for the vast majority of skin cancer deaths (American-Cancer-Society, 2014). According to latest reports, melanoma causes over 20,000 deaths annually in Europe (Forsea et al., 2012). In 2014, the American Cancer Society also reported that the number of new diagnosed cases is 76,100 with 9710 estimated deaths (American-Cancer-Society, 2014). Nevertheless, melanoma is the most treatable kind of cancer if diagnosed early.

Melanoma is clinically diagnosed through visual inspection and deep analysis of the lesion, using clinical imaging techniques such as dermoscopic imaging. The clinical diagnosis of early stage melanoma is commonly based on the “ABCDE” rule (Abbasi et al., 2004), defined as Asymmetry, irregular Borders, variegated Colors, Diameters greater than 6 mm and its Evolution over time. These inspections and analysis are challenging since, first different lesions such as melanoma and dysplastic nevi share similar characteristics in terms of “ABCD” and second the necessity to perform patient follow-up over the years. Therefore, the research communities have dedicated their efforts to develop computerized lesion analysis algorithms for classification of melanoma lesions.

When studying skin lesions, akin to other medical applications, the percentage of malignant cases is far less when compared with benign cases. This problem is frequently referred as “class imbalance” problem (Prati et al., 2009) and has been encountered in multiple areas such as telecommunication managements, bioinformatics, fraud detection, and medical diagnosis. Imbalanced data substantially compromises the learning process since most of the standard machine learning algorithms expect balanced class distribution or an equal misclassification cost (He et al., 2009).

Medical data are prone to such drawbacks due to the fact that the portion of diseased samples or patients is far lower than healthy cases. Furthermore, the detection and classification of minority malignant cases are highly essential so that the Sensitivity (SE) of developed algorithms need to be maximized. Consequently, the problem of imbalanced data is usually addressed by employing different techniques which do not impair the topology of the data. Despite the fact that classification of malignant melanoma has been extensively studied (****, a), up to our knowledge, only few works tackled the issue implied by imbalanced dataset (Barata et al., 2014, Celebi et al., 2007). Barata *et al.* generate new synthetic samples by adding a Gaussian noise with fixed parameters to the samples belonging to



(a) Melanoma lesion (b) Dysplastic lesion (c) Benign lesion

Figure 1: Samples of PH^2 dataset, representing melanoma, dysplastic and benign lesions, respectively.

the minority class (Barata et al., 2014). Celebi *et al.* and Capdehourat *et al.* over-sampled their dataset using Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002) to improve the SE of their algorithm (Celebi et al., 2007, Capdehourat et al., 2009).

This paper provides an insight to the specific problem of classification of imbalanced dataset for malenoma. To proceed, we review different techniques proposed by the machine learning community and compile a comprehensive quantitative evaluation. The rest of this paper is organized as follows: an overview of the classification framework designed to investigate data balancing techniques is presented in Sect. 2. The balancing strategies are explained in depth in Sect. 3 and the validation and classification are discussed in Sect. 4. A quantitative evaluation is discussed in Sect. 5 followed by a concluding section.

2 MATERIAL AND METHODS

Figure 4 illustrates and summarizes the experiment designed to explore the data imbalance problem during the classification of dermoscopic images. The experimentation is based on the works presented in (****, a, ****, b) and follows a cross-validated classification evaluation framework. Details of the dataset used for the experiments are given in Sect. 2.1. The extracted features correspond to the highest performing subset of features according to the latter mentioned studies and are presented in Sect. 2.2.

2.1 Dataset

In order to allow future comparisons, we choose to work with the only public dermoscopic dataset PH^2 (Barata et al., 2014). This dataset is acquired at *Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal* (Barata et al., 2014) with Tuebinger Mole Analyzer system with a magnification of 20×. The 8-bits RGB color dermoscopic images were obtained under the same conditions with a resolution of 768 px × 560 px. This dataset contains 200 dermoscopic images divided into 160 benign and dysplastic and 40 melanoma lesions. In terms of Ground Truth (GT), histological diagnosis and segmentation of the lesions are provided.

Due to an imbalance limitation of one of the techniques here studied, the experimentation is conducted on a data subset with an imbalance ratio of 1:3. Thus, the subset is composed of 39 melanoma and 117 benign and dysplastic lesions, randomly selected. Figure 1 shows three samples of this dataset, representing melanoma, dysplastic, and benign lesion, respectively.

2.2 Feature extraction

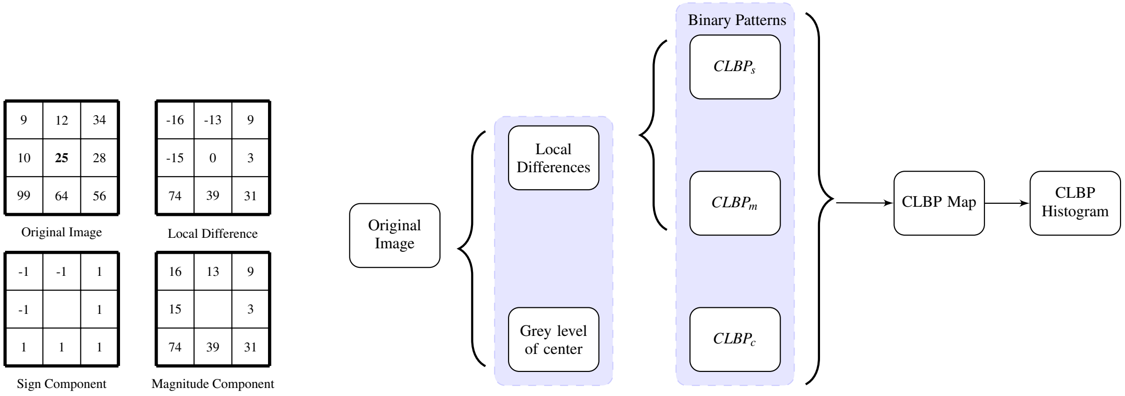
The color variance and histogram (C_1) descriptor contains the mean and variance of the color channels {R, G, B, H, S, V, L, A, B} and a 42 bins histogram for each channel of the set {R, G, B}. Thus, the final descriptor is made of 144 features.

The opponent color space angle and hue histogram (C_2) is a robust and rotation invariant feature descriptor derived from the RGB channels (Van De Weijer and Schmid, 2006), such that:

$$H = \arctan\left(\frac{\sqrt{3}\,(R-G)}{R+G-2B}\right) ,$$

$$\theta_d^O = \arctan\left(\frac{\sqrt{3}\,(R'_d-G'_d)}{R'_d+G'_d-2B'_d}\right) , \qquad (1)$$

where d denotes the spatial coordinates of (x,y) and R'_d, G'_d, B'_d denote the first order derivatives of RGB channels with respect to the coordinates. This color descriptor is built by taking a 42 bins histogram for the opponent angle θ_d^O and the hue channel (H), for a final descriptor size of 84 dimensions.



(a) (b)

Figure 2: CLBP descriptor process, (a) represents an example on how local distances, sign and magnitude components are calculated and (b) shows an overall view of CLBP process.

Completed Local Binary Pattern (CLBP) (T_1) is a completed modeling of Local Binary Pattern, especially designed for texture classification (Guo and Zhang, 2010). This descriptor encodes the magnitude and sign differences of the central pixel with its neighbors and the grey level of the central points in the local patterns rather than only the sign differences (see Fig 2). The sign $CLBP_S$, magnitude $CLBP_M$, and central grey level $CLBP_C$ binary pattern are created by encoding the local distance components and the central grey levels to binary patterns. The CLBP are calculated for each pixel in a given image and the final descriptor is defined as their histogram. The rotation invariant, uniform, and normalized CLBP features is calculated considering a radius of 24 px. The descriptor is composed of 26 dimensions.

Gabor filter (T_2) is a linear filter which is defined as a modulation of a Gaussian kernel with a sinusoidal wave. This filter is formulated in Eq. (2) as two Gaussians with standard deviations of σ_x and σ_y that vary along x and y axes and it is modulated by a complex sinusoidal with a wavelength of λ. Here θ represents the orientation of the Gabor filter, ψ is the phase offset and s is the scale factor. The filter bank is created using six different orientations equally spaced in the interval $[0,\pi]$, along 4 scales with a downsizing factor of 2:

$$g(x,y) = \exp\left(-\left(\frac{x'^2}{2\sigma_x^2}+\frac{y'^2}{2\sigma_y^2}\right)\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) , \qquad (2)$$

where

$$x' = s\,(x\cos\theta + y\sin\theta) ,$$
$$y' = s\,(-x\sin\theta + y\cos\theta) .$$

The final descriptor is composed 48 feature dimensions.

3 BALANCING STRATEGIES

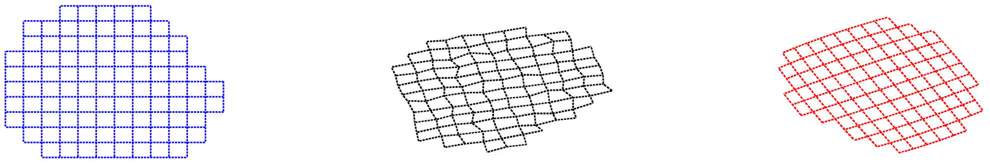
Considering a binary classification problem, the class with the smallest number of samples is defined as the *minority* class and its counterpart is defined as the *majority* class. The problem of data balancing corresponds to equalize the number of samples of both the minority and majority classes. This task can be achieved in either data or feature space.

3.1 Data space sampling

Data space sampling is related with the generation of new synthetic samples by modifying the original data ahead of any feature extraction processes. Over-Sampling (OS) is performed on the original dataset by generating synthetic melanoma images based on two types of deformation (****, b). Furthermore, cubic b-spline interpolation is used with both methods to approximate non-integer points in the image. These deformations are considered since they are more likely to occur, due to non-planar surface of some body parts, skin wrinkles, camera rotation, and position.

Random Deformation using Gaussian Motion achieved by deforming the original image by adding a random Gaussian motion $\mathcal{N}(\mu,\sigma) = (0,5)$ at each pixel compounded with a global rotation of 80°.

Barrel Deformation corresponds to a deformation of the original image using barrel distortion compounded with a global rotation of 145°.



(a) (b) (c)

Figure 3: Data space transformation: (a) original synthetic data, (b) RDGM deformation, (c) BD deformation.

A synthetic example illustrating the results of these deformation is presented in Fig. 3.

3.2 Feature space sampling

Three strategies can be employed to overcome the problem of imbalanced dataset: (i) Under-Sampling (US), (ii) OS, and (iii) a combination of both. The following sections give an overview of the techniques used to tackle this issue.

3.2.1 Under-Sampling

Considering the problem of imbalanced, US is performed such that the number of samples of the majority class is reduced to be equal to the number of samples of the minority class. The following methods are considered to perform such balancing.

Random Under-Sampling (RUS) is performed by randomly selecting without replacement a subset of samples from the majority class such that the number of samples is then equal in both minority and majority classes.

Tomek Link (TL) can be used to under-sample the majority class of the original dataset (Tomek, 1976). Let define a pair of Nearest Neighbour (NN) samples (x_i, x_j) such that their associated class label $y_i \neq y_j$. The pair (x_i, x_j) is defined as a TL if, by relaxing the class label differentiation constraint, there is no other sample x_k defined as the NN of either x_i or x_j. US is performed by removing the samples belonging to the majority class and forming a TL. It can be noted that this US strategy does not enforce a strict balancing between the majority and the minority classes.

Clustering Under-Sampling (CUS) refers to the use of a k-means to cluster the feature space such that k is set to be equal to the number of samples composing the minority class. Hence, the centroids of these clusters define the new samples of the majority class.

NearMiss offers three different methods to under-sample the majority class (Mani and Zhang, 2003). In NearMiss-1 (NM1), samples from the majority class are selected such that for each sample, the average distance to the k NN samples from the minority class is minimum. NearMiss-2 (NM2) diverges from NM1 by considering the k farthest neighbours samples from the minority class. In NearMiss-3 (NM3), a subset M containing samples from the majority class is generated by finding the m NN from each sample of the minority class. Then, samples from the subset M are selected such that for each sample, the average distance to the k NN samples from the minority class is maximum. In our experiment, k and m are fixed to 3.

Neighborhood Cleaning Rule (NCR) consists of applying two rules depending on the class of each sample (Laurikkala, 2001). Let define x_i as a sample of the dataset with its associated class label y_i. Let define y_m as the class of the majority vote of the k NN of the sample x_i. If y_i corresponds to the majority class and $y_i \neq y_m$, x_i is rejected from the final subset. If y_i corresponds to the minority class and and $y_i \neq y_m$, then the k NN are rejected from the final subset. In our experiment k is fixed to 3.

3.2.2 Over-Sampling

In the contrary, the data balancing can be performed by OS in which the new samples belonging to the minority class are generated aiming at equalizing the number of samples in both classes. Two different methods are considered.

Random Over-Sampling (ROS) is performed by randomly replicating the samples of the minority class such that the number of samples is equal in both minority and majority classes.

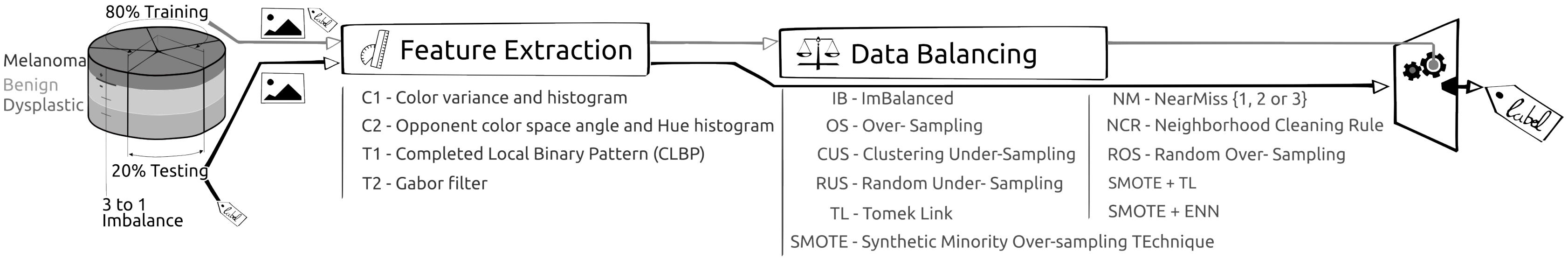

80% Training
Melanoma
Benign
Dysplastic
20% Testing
3 to 1
Imbalance
Feature Extraction
C1 - Color variance and histogram
C2 - Opponent color space angle and Hue histogram
T1 - Completed Local Binary Pattern (CLBP)
T2 - Gabor filter
Data Balancing
IB - ImBalanced
OS - Over- Sampling
CUS - Clustering Under-Sampling
RUS - Random Under- Sampling
TL - Tomek Link
SMOTE - Synthetic Minority Over-sampling TEchnique
NM - NearMiss {1, 2 or 3}
NCR - Neighborhood Cleaning Rule
ROS - Random Over- Sampling
SMOTE + TL
SMOTE + ENN
label

Figure 4: Framework outline

Table 1: The obtained results with different balancing techniques for color and texture features using a RF classifier. The first and second highest results for each feature set are highlighted in dark and lighter gray colors, respectively.

Features	Color						Texture						Combined					
	C_1		C_2		$C_{1,2}$		T_1		T_2		$T_{1,2}$		$T_1, C_{1,2}$		$T_2, C_{1,2}$		$T_{1,2}, C_{1,2}$	
Balancing techniques	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP
IB	52.5	89.6	75.0	88.7	71.2	87.5	38.7	91.7	60.0	96.2	66.2	93.7	73.7	89.6	71.2	89.6	71.2	92.5
OS	93.7	66.7	80.0	86.2	82.5	87.1	43.7	83.7	72.5	90.0	70.0	91.7	77.5	87.1	81.2	88.3	78.7	88.3
ROS	55.0	80.8	80.0	84.2	72.5	85.4	42.5	82.1	60.0	89.2	66.2	87.9	75.0	85.4	73.7	86.2	73.7	85.8
SMOTE	60.0	82.5	78.7	84.6	75.0	70.0	56.2	74.2	61.2	87.5	84.2	87.1	78.7	85.0	73.7	84.6	73.7	85.0
RUS	72.5	72.9	86.2	80.0	78.7	80.0	67.5	53.3	76.2	76.2	85.0	78.7	91.2	75.0	85.0	78.7	92.5	78.3
TL	51.2	86.2	76.2	87.9	67.5	88.3	37.5	87.9	65.0	90.4	68.7	91.7	73.7	88.7	63.7	90.0	72.5	91.2
CUS	81.2	67.9	80.0	84.6	86.2	80.4	56.2	65.8	70.0	77.5	85.0	77.1	83.7	81.2	80.0	84.2	83.7	82.9
NM1	67.5	72.1	86.2	79.2	85.0	82.5	72.5	43.7	80.0	62.5	87.5	66.7	85.0	82.1	86.2	80.4	87.5	80.8
NM2	70.0	72.9	86.2	81.2	85.0	82.9	76.2	48.7	86.2	40.8	86.2	51.2	87.5	82.1	92.5	77.5	91.2	81.7
NM3	82.5	75.0	87.5	80.8	85.0	80.4	73.7	55.8	72.5	82.5	82.5	80.4	83.7	81.2	85.0	80.0	86.2	80.4
NCR	66.2	76.7	87.5	81.2	85.0	82.1	67.5	67.9	75.0	85.8	82.5	83.3	86.2	81.7	82.5	85.0	83.7	85.4
SMOTE + ENN	76.2	73.3	85.0	81.2	85.0	82.1	81.2	56.2	76.2	82.1	80.0	79.6	86.2	81.2	83.7	82.5	78.7	82.9
SMOTE + TL	75.0	73.7	83.7	82.5	87.5	80.8	72.5	59.2	77.5	82.1	78.7	78.7	85.0	82.1	77.5	82.9	88.7	82.5

SMOTE is a method to generate synthetic samples in the feature space (Chawla et al., 2002). Let define x_i as a sample belonging to the minority class. Let define x_{nn} as a randomly selected sample from the k NN of x_i, with k set to 3. Therefore, a new sample x_j is generated such that $x_j = x_i + \sigma(x_{nn} - x_i)$, where σ is a random number in the interval $[0,1]$.

3.2.3 Combination of OS and US

Subsequently, OS methods can be combined with US methods to clean the subset created. In that regard, two different combinations are tested.

SMOTE + TL are combined to clean the samples created using SMOTE (Batista et al., 2003). SMOTE over-sampling can lead to overfitting which can be avoided by removing the TL from both majority and minority classes (Prati et al., 2009).

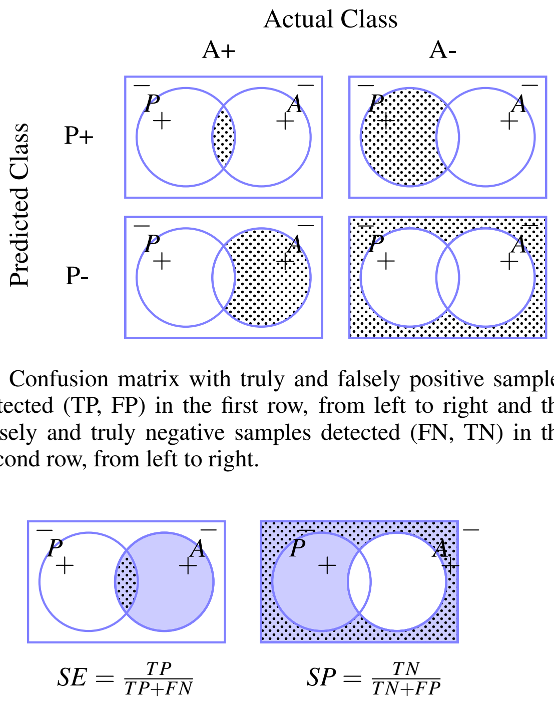
SMOTE + Edited Nearest Neighbour are combined for the same aforementioned reason (Batista et al., 2004).

4 CLASSIFICATION

The classification is performed using a Random Forests (RF) classifier. RF is an ensemble of decision trees (Breiman, 2001) which generalizes the classification process by applying two types of randomization: at the tree level, each tree is fed by a bootstrap made of S' samples built from the original data of size S such that $S = S'$, and at the node level, a subset of feature dimensions m is randomly selected from the original dimension M such that $m = \sqrt{M}$. The trees in RF are grown to their maximum length without any pruning. Each tree in the ensemble casts a unit vote in the final prediction and the final prediction is based on combination of all the votes. RF is used with 100 un-pruned trees and the original feature dimension of size $M = \{144, 84, 228, 26, 48, 74, 254, 276, 302\}$

4.1 Validation

We used a 10-fold cross-validation scheme to validate our classifier with stratified sampling. However, differently from the usual 10-fold cross-validation, 8 folds were kept for training and 2 folds for testing at each iteration. The training set is balanced using previously described imbalanced techniques. The classification performance are reported in terms of average SE (TPR) and Specificity (SP) (TNR) over 10 runs of cross-validation. The visual and analytic interpretation of these evaluation measures are depicted in Fig. 5.



(a) Confusion matrix with truly and falsely positive samples detected (TP, FP) in the first row, from left to right and the falsely and truly negative samples detected (FN, TN) in the second row, from left to right.

(b) Sensitivity and Specificity evaluation, corresponding to the ratio of the doted area over the blue area.

Figure 5: Evaluation metrics: (a) confusion matrix, (b) Sensitivity - Specificity

To select the best performance, similarly to (Barata et al., 2013) we consider to evaluate the results based on a cost function, which defines the trade off between SE and SP. This function is formulated as:

$$C = \frac{c_{10}(1-SE) + c_{01}(1-SP)}{c_{10} + c_{01}}\,, \qquad (3)$$

where, c_{10} and c_{01} are the costs of incorrectly classifying a melanoma and non-melanoma lesions, respectively. In cancer classification such as melanoma, correctly identifying the cancer lesions has high importance (i.e., high SE). Thus incorrect classification of melanoma is not desired and c_{10} is a greater error and evidently more costly and should be penalized more. In order to achieve a high SE without significantly reducing the value of SP, Barata *et al.* proposed to set $c_{10} = 1.5 \times c_{01}$ and $c_{01} = 1$ (Barata et al., 2013). We considered the same configuration for our cost function.

Table 2: The classification costs, C for different balancing techniques and feature sets.

Balancing techniques	Classification cost, C								
	C_1	C_2	$C_{1,2}$	T_1	T_2	$T_{1,2}$	$T_1,C_{1,2}$	$T_2,C_{1,2}$	$T_{1,2},C_{1,2}$
IB	0.3267	0.1950	0.2225	0.4008	0.2550	0.2275	0.1992	0.2142	0.2025
OS	0.1708	0.1750	0.1567	0.4025	0.2050	0.2133	0.1867	0.1592	0.1742
OS	0.3467	0.1833	0.2233	0.4167	0.2833	0.2508	0.2083	0.2125	0.2142
SMOTE	0.3100	0.1892	0.2133	0.3658	0.2825	0.2317	0.1875	0.2192	0.2175
RUS	0.2733	0.1625	0.2075	0.3817	0.2375	0.1750	0.1525	0.1750	0.1317
TL	0.3475	0.1908	0.2417	0.4233	0.2483	0.2208	0.2025	0.2575	0.2000
CUS	0.2408	0.1817	0.1608	0.3992	0.2700	0.1817	0.1725	0.1833	0.1658
NM1	0.3067	0.1658	0.1600	0.3900	0.2700	0.2083	0.1617	0.1608	0.1517
NM2	0.2883	0.1575	0.1583	0.3475	0.3192	0.2775	0.1467	0.1350	0.1258
NM3	0.2050	0.1517	0.1683	0.3342	0.2350	0.1833	0.1725	0.1700	0.1608
NCR	0.2958	0.1500	0.1617	0.3233	0.2067	0.1717	0.1558	0.1650	0.1558
SMOTE+ENN	0.2492	0.1650	0.1617	0.2875	0.2142	0.2017	0.1575	0.1675	0.1958
SMOTE+TL	0.2550	0.1675	0.1517	0.3283	0.2067	0.2125	0.1617	0.2033	0.1375

5 EXPERIMENTAL RESULTS

The classification results are reported in Table 1 using the aforementioned features, the RF classifier, and the different imbalancing techniques presented in Sect. 3 and Sect. 4.

Table 1 can be divided into three main parts representing the results using imbalance data (IB), the balancing in the data space OS and the balancing in the feature space. These strategies are separated by a double horizontal line. The strategies performed in the feature space are subdivided into either OS or US or a combination of OS follow by US (see horizontal dashed line in Table 1).

In this table, based on the previously defined cost function, the best performance for each feature set are highlighted in the shaded cells. Table 2 shows the obtained cost value for each configuration. Strategies with low cost function are synonymous with a better SE and SP trade-off.

The obtained results indicate that balancing techniques are essential and improve the classification performance. For this case study the US techniques and their combination with OS techniques outperform the OS techniques. Due to the characteristics similarities of melanoma and dysplastic lesions, it is expected to have correlated feature space among melanoma and dysplastic lesions. Subsequently, the miss-leading samples could be removed using US and lead to better performance. Specifically to our purpose, NM2 algorithm with the combination of all the features ($T_{1,2}C_{1,2}$) with the lowest cost value, achieve the highest SE and SP of 91.2% and 81.7%. Using the same feature combination, RUS achieve the second lowest cost with SE and SP of 92.5% and 78.3%, respectively. The NM2 algorithm also achieve the third lowest cost with combination of Gabor and color features ($T_2,C_{1,2}$) with SE and SP of 92.5% and 77.5%, respectively. Focusing only on OS techniques, OS in data space outperforms the techniques performing in feature space.

Comparing the color features, opponent color angle and hue histogram feature descriptor, C_2, has a better performance than well-used color statistics, C_1. In texture domain, Gabor descriptor, T_2, outperforms CLBP features, T_1. Finally, the combination of color and texture features outperforms any other feature combination.

6 CONCLUSION

In this paper, we analyzed the impact of data balancing techniques for the classification of malignant melanoma. Therefore, we presented an extensive comparison of twelve OS and US techniques in both feature and data space. These techniques were evaluated on the only public dermoscopic dataset, the PH^2 dataset, in order to provide a chance for future comparison. The obtained results particularly highlight the advantage of balancing the training set over using the original data, particularly for the methods based on US (NM2,NCR) and combination of OS and US in feature space. Furthermore, OS in data space outperforms the techniques performing in the feature space. This study also showed that combining color and texture features will lead to better performance.

REFERENCES

**** (****a). ********************. ********************, **:******.

**** (****b). ********************. In ********************, page ***********. ********************.

Abbasi, N. R., Shaw, H. M., et al. (2004). Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *Jama*, 292(22):2771–2776.

American-Cancer-Society (2014). *Cancer facts & figures 2014*.

Barata, C., Marques, J. S., and Emre Celebi, M. (2013). Towards an automatic bag-of-features model for the classification of dermoscopy images: The influence of segmentation. In *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on*, pages 274–279. IEEE.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., and Marques, J. (2014). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal,*, 8(3):965–979.

Batista, G. E., Bazzan, A. L., and Monard, M. C. (2003). Balancing training data for automated annotation of keywords: a case study. In *WOB*, pages 10–18.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Capdehourat, G., Corez, A., Bazzano, A., and Musé, P. (2009). Pigmented skin lesions classification using dermatoscopic images. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 537–544. Springer.

Celebi, M. E., Kingravi, H. A., et al. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, pages 321–357.

Forsea, A., Del Marmol, V., de Vries, E., Bailey, E., and Geller, A. (2012). Melanoma incidence and mortality in europe: new estimates, persistent disparities. *British Journal of Dermatology*, 167(5):1124–1130.

Guo, Z. and Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663.

He, H., Garcia, E., et al. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284.

Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution*. Springer.

Mani, I. and Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Prati, R. C., Batista, G. E., and Monard, M. C. (2009). Data mining with imbalanced class distributions: concepts and methods. In *IICAI*, pages 359–376.

Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.*, 6:769–772.

Van De Weijer, J. and Schmid, C. (2006). Coloring local feature extraction. In *Computer Vision–ECCV 2006*, pages 334–348. Springer.