Tackling the Curse 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. 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) outperforms other methods by achieving Sensitivity (SE) and Specificity (SP) of 92.50 % and 77.50 %, respectively. More generally, the reported results highlight that methods based on OS in data space 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 (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 (Society, 2014). Nevertheless, melanoma is the most treatable kind of cancer if diagnosed early.

The clinical diagnosis of early stage melanoma is commonly based on the "ABCDE" rule (Abbasi et al., 2004), defined as Asymmetry, irregular Borders, variegated Colours, Diameters greater than 6 mm and Evolving stages over time. In addition, melanoma are clinically diagnosed through visual inspection and deep analysis of the lesion, using clinical imaging techniques such as dermoscopic imaging. However, these inspections and analysis are challenging due to the similarity of the different lesion types (dysplastic and melanoma) and the necessity to follow-up patient over years. Therefore, the research communities have dedicated their efforts to develop computerized lesion analysis algorithms for classification of melanoma lesions. However, akin to other medical applications, the percentage of melanoma cases in comparison with benign and dysplatic cases is far less. This problem is frequently referred as "class imbalanced" 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 vitiate the topology of the data. Despite the fact that classification of malignant melanoma has been extensively studied (****, a), up to our knowledge, only two 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 the minority class (Barata et al., 2014). Celebi *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).

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 while these strategies are described in Sect. 3. A quantitative evaluation is discussed in Sect. 5 followed by a concluding section.

2 MATERIAL AND METHODS

Figure 3 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 summarized in Sect. 2.2. The balancing strategies are explained in great details in Sect. 3 and finally the validation and classification is discussed in Sect. **??**.

2.1 Dataset

The PH^2 dermoscopic dataset which is acquired at *Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal* is used (Barata et al., 2014). The dermoscopic dataset is acquired with Tuebinger Mole Analyzer system with a magnification of $20\times$. The 8-bits RGB color dermoscopic images were obtained under the same conditions with a resolution of $768\,\text{px} \times 560\,\text{px}$. This dataset contains 200 dermoscopic images divided into 160 benign and dysplastic and 40 melanoma lesions. Moreover, each lesion is segmented and histological diagnosis are provided. In this study, we conducted the experiments with a subset of 39 melanoma and 117 benign and dysplastic lesions with an imbalance ratio of 1:3.

2.2 Feature extraction

Color variance and histogram (C_1) descriptors contain the mean and variance of the nine channels (R, G, B, H, S, V, L, A, B) and the histogram of the R, G and B channels.

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):

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

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. The color descriptor is built by taking histogram of the opponent angle θ_d^O and the hue channel (H).

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 in the local patterns rather than only the sign differences. The CLBP are calculated for each pixel in a given image and their histogram defines the final descriptor.

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 Gaussian 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:

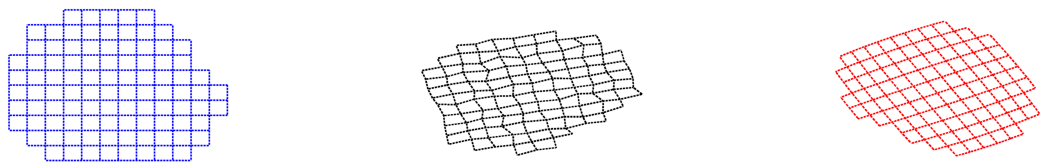
$$g(x,y) = \exp\left(-\left(\frac{x'^2}{2\sigma_x^2}+\frac{y'^2}{2\sigma_y^2}\right)\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) , \qquad (2)$$

where

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

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.



(a) (b) (c)

Figure 1: Data space transformation: (a) original synthetic data, (b) RDGM deformation, (c) BD deformation.

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.

Random Deformation using Gaussian Motion achieved by deforming the original image by adding a random Gaussian motion $\mathcal{N}(\mu,\sigma) = (5,5)$ 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 synthetic example illustrating the results of these deformation is presented in Fig. 1.

3.2 Feature space sampling

Considering the problem of imbalanced, Under-Sampling (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 (NM) 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 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.

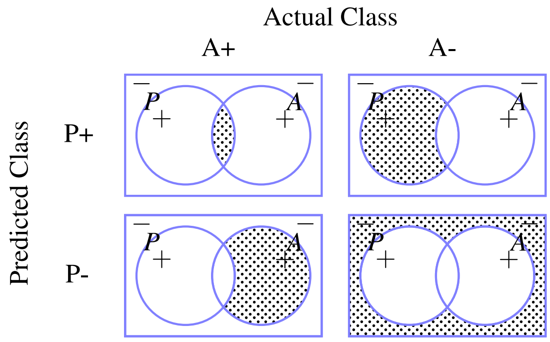
Random Over-Sampling (ROS) is performed by randomly replicating the samples of the minority class such that the number of samples is equal in both minority and majority classes.

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. 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 $[0, 1]$.

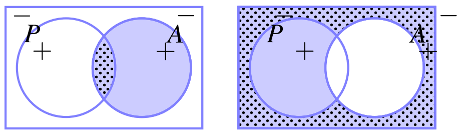
Subsequently, OS methods can be combined with US methods to clean the subset created. In that regard, two different combinations are tested.

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

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



(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.



$SE = \frac{TP}{TP+FN}$ $\quad$ $SP = \frac{TN}{TN+FP}$

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

Figure 2: Evaluation metrics: (a) confusion matrix, (b) Sensitivity - Specificity

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 using different bootstrap samples of the original data and splitting the feature dimensions at each node. Each bootstrap with M attributes is used to train one decision tree and at each node in the tree, the best decision is taken based on gini criterion on the randomly selected m attributes (such as $m << M$). 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.

4.1 Validation

The validation model used is a 10-fold cross-validation in which 80 % of the data are used for training and 20 % are used for testing. 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. 2.

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. 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). The two highest SE for each feature set are highlighted in dark and light gray cell colors, respectively.

The obtained results indicate that balancing techniques are essential and improve the classification performance. However, the improvements in comparison to imbalanced classification is evident. For this case study the US 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 is the algorithm maximizing the sensitivity and in overall, NM algorithms perform the best on our dataset. However, NCR algorithm (see results highlighted in blue in Table 1) achieves the best performance, considering a trade-off between SE and SP. Focusing only on OS techniques, OS in data space outperforms the techniques performing in feature space.

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 a subset of PH^2 dataset with an imbalanced ration of 1:3. The obtained results particularly highlight the advantage of balancing the training set over using the original data, particularly for the methods based on OS in data space and US in feature space (NM).

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.

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.

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.

Society, A. C. (2014). Cancer facts & figures 2014.

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.

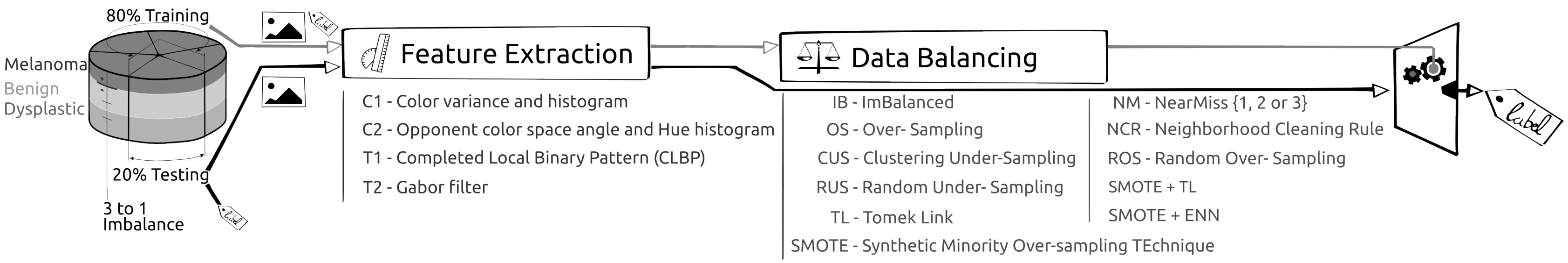


Figure 3: Framework outline

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	52.50	89.58	75.00	88.75	71.25	87.50	38.75	91.67	60.00	96.25	66.25	93.75	73.75	89.58	71.25	89.58	71.25	92.50
OS	93.75	66.67	80.00	86.25	82.50	87.08	43.75	83.75	72.50	90.00	70.00	91.67	77.50	87.08	81.25	88.33	78.75	88.33
ROS	55.00	80.83	80.00	84.17	72.50	85.42	42.50	82.08	60.00	89.17	66.25	87.92	75.00	85.42	73.75	86.25	73.75	85.83
SMOTE	60.00	82.50	78.75	84.58	75.00	70.00	56.25	74.17	61.25	87.50	84.17	87.08	78.75	85.00	73.75	84.58	73.75	85.00
RUS	72.50	72.92	86.25	80.00	78.75	80.00	67.50	53.33	76.25	76.25	85.00	78.75	91.25	75.00	85.00	78.75	92.50	78.33
TL	51.25	86.25	76.25	87.92	67.50	88.33	37.50	87.92	65.00	90.42	68.75	91.67	73.75	88.75	63.75	90.00	72.50	91.25
CUS	81.25	67.92	80.00	84.58	86.25	80.42	56.25	65.83	70.00	77.50	85.00	77.08	83.75	81.25	80.00	84.17	83.75	82.92
NM1	67.50	72.08	86.25	79.17	85.00	82.50	72.50	43.75	80.00	62.50	87.50	66.67	85.00	82.08	86.25	80.42	87.50	80.83
NM2	70.00	72.92	86.25	81.25	85.00	82.92	76.25	48.75	86.25	40.83	86.25	51.25	87.50	82.08	92.50	77.50	91.25	81.67
NM3	82.50	75.00	87.50	80.83	85.00	80.42	73.75	55.83	72.50	82.50	82.50	80.42	83.75	81.25	85.00	80.00	86.25	80.42
NCR	66.25	76.67	87.50	81.25	85.00	82.08	67.50	67.92	75.00	85.83	82.50	83.33	86.25	81.67	82.50	85.00	83.75	85.42
SMOTE + ENN	76.25	73.33	85.00	81.25	85.00	82.08	81.25	56.25	76.25	82.08	80.00	79.58	86.25	81.25	83.75	82.50	78.75	82.92
SMOTE + TL	75.00	73.75	83.75	82.50	87.50	80.83	72.50	59.17	77.50	82.08	78.75	78.75	85.00	82.08	77.50	82.92	88.75	82.50