Building of Readable Decision Trees for Automated Melanoma Discrimination

Keiichi Ohki[1], M. Emre Celebi[2], Gerald Schaefer[3], and Hitoshi Iyatomi[4(✉)]

[1] Graduate School of Science and Engineering, Hosei University, Saitama, Japan
[2] Department of Computer Science, Louisiana State University in Shreveport, Shreveport, USA
[3] Department of Computer Science, Loughborough University, Loughborough, UK
[4] Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
iyatomi@hosei.ac.jp

Abstract. Even expert dermatologists cannot easily diagnose a melanoma, because its appearance is often similar to that of a nevus, in particular in its early stage. For this reason, studies of automated melanoma discrimination using image analysis have been conducted. However, no systematic studies exist that offer grounds for the discrimination result in a readable form. In this paper, we propose an automated melanoma discrimination system that it is capable of providing not only the discrimination results but also their grounds by means of utilizing a Random Forest (RF) technique. Our system was constructed based on a total of 1,148 dermoscopy images (168 melanomas and 980 nevi) and uses only their color features in order to ensure the readability of the grounds for the discrimination results. By virtue of our efficient feature selection procedure, our system provides accurate discrimination results (a sensitivity of 79.8 % and a specificity of 80.7 % with 10-fold cross-validation) under human-oriented limitations and presents the grounds for the results in an intelligible format.

1 Introduction

Melanoma is a fatal skin cancer and among the cancer types that very frequently become metastatic. The five-year survival rate of patients diagnosed with advanced melanoma is reported to be about 7 % [1]. Therefore, early detection of a melanoma is crucial; however, even experts cannot easily diagnose it, in particular in its early stage. Dermoscopy [2], a non-invasive skin imaging technique, and associated diagnosis criteria, such as the ABCD rules [3] and 7-point check list [4], were developed to improve diagnostic accuracy. However, dermoscopic diagnosis is often subjective and is therefore associated with poor reproducibility and low accuracy, in particular when performed by inexperienced dermatologists. Argentiano et al. reported that the accuracy of melanoma diagnosis performed by experts using the criteria mentioned above was about 75–84 % and both the inter-observer and intra-observer diagnosis agreement were large [5]. If an accurate, easy, and objective diagnosis method could be established,

G. Bebis et al. (Eds.): ISVC 2015, Part II, LNCS 9475, pp. 712–721, 2015.
DOI: 10.1007/978-3-319-27863-6_67

it would facilitate early detection of melanomas and reduce melanoma-related deaths. To address this issue, studies on automated melanoma diagnosis have been conducted [6,10]. Rubegni et al. [6] trained an artificial neural network (ANN) using 200 melanomas and 350 nevi and 48 extracted parameters. Their ANN achieved a sensitivity (SE: melanoma detection accuracy) of 94.3 % and a specificity (SP: nevus detection accuracy) of 93.8 % on a test dataset consisting of 17 melanomas and 21 nevi. Celebi et al. [7] developed a support vector machine classification model using 88 melanomas and 479 nevi. Their system achieved an SE of 93.3 % and SP of 92.3 % under Monte Carlo cross-validation. In 2004, we opened our Internet-based melanoma screening system for dermatologists [8,9]. Our current system (http://dermoscopy.k.hosei.ac.jp) comprises approximately 1,500 dermoscopy images and achieved both an SE and SP of 86 % for tumors in normal body parts [9], and an SE of 91 % and SP 93 % for tumors in acral volar regions [10] under leave-one-out cross-validation. Our system provides discrimination results as well as a malignancy score. Although several conventional systems attained a discrimination performance equivalent or superior to that of expert dermatologists in numeric terms, most of these automated systems used a limited type of tumor image, the condition of which was good, i.e., the brightness and color were appropriate and many artifacts, such as hairs, bubbles, and the border of the scope, were not included. Accordingly, taking the practical use of these automated systems into consideration, it is considered desirable that the system can offer the grounds for the discrimination results, i.e., the reason(s) why the system yields the results, in order to ensure reliability. From this viewpoint, previously we proposed a quantification method for each clinical finding defined in the ABCD rule [3] and 7-point checklist [4], which attained a good estimation performance [11]. However, this research was conducted based on only 105 dermoscopy images because of the difficulty in obtaining a reliable gold standard (15 items for each case; diagnosis by multiple experts was required), and therefore, further investigations were required in order to attain robustness. In this paper, we propose an automated melanoma discrimination system utilizing a Random Forest (RF) technique that is capable of providing discrimination and its associated grounds [12]. RF is a machine-learning algorithm that aggregates several decision "trees" as weak learners to form a highly accurate classifier "forest." The discrimination result of each decision tree is obtained by following a path from the root node to the leaf node and the majority vote of the nodes determines the final decision of the RF. In this study, the grounds for the discrimination results is expressed in the form of the built tree, which has important image features. So that the grounds for the discrimination results are readable, the image features used in the RF classifier should be intelligible and the number of tree layers limited. To achieve this, we have to seek important and readable image features for classifying melanomas that attain a certain classification accuracy under these limitations. In this paper, we also propose an effective feature selection method for constructing a sophisticated RF classifier.

2 Random Forests

RF is an accurate classification algorithm that aggregates a multitude of decision trees as weak learners. Each tree is built based on a limited number of randomly selected samples and their features, allowing them to overlap. This strategy results in a lower correlation among each tree and accordingly, the formed "forest" is expected to be resistant to over-fitting and is capable of producing an accurate classification.

In each node in a decision tree, the feature used for dividing data was determined by means of its entropy. When we assume that data are divided into N groups at the branch node k, the entropy of the k-th node $H(k,p)$ is expressed as

$$H(k,p) = -\sum_{i=1}^{N} p(C_i|p)\log p(C_i|p), \tag{1}$$

where $p(C_i|p)$ represents the ratio according to which given data are divided into class $C_i (i = 1, 2, ..., N)$ using feature p. Here, an expected entropy reduction at the node k is

$$\eta(k,p) = H(k,p) - \sum_{j=1}^{M(k)} H(k_j,p), \tag{2}$$

where $k_j (j = 1, 2, ..., M(k))$ is the child node of the node k. This amount, $\eta(k,p)$, represents the effect of the data partitioning by feature p at the node k. In the k-th node, the applicable feature $p^*(k)$ is selected from $\forall\ p$ such that the corresponding $\eta(k,p)$ is a maximum.

$$p^*(k) = \underset{p}{\operatorname{argmax}}\, \eta(k,p). \tag{3}$$

Note that the applicable number of feature candidates for each node, i.e., the parameter pool, is suggested to be the square root of that of the all features in order to ensure independence among the trees [12].

3 Melanoma Discrimination System Capable of Offering the Grounds

Our proposed system is composed of the following five phases.

- Tumor area extraction
- Calculation of the image features
- Selection of the important features with RF
- Training of RF
- Presentation of the classification results and decision tree.

The first to the fourth phases are for constructing the RF classifier and the final phase yields the discrimination results.

In pattern recognition problems, the design and selection of important features for recognition constitute one of the most important processes. RF plants a multitude of highly independent decision trees by sampling the training data and their features randomly. Here, when the training data contain considerable noise and/or the correlations among their features are high, the overall performance of the RF is decreased. Accordingly, we also propose a novel feature selection method using RF, the effectiveness of which we evaluated. The details are described in the following sections.

3.1 Tumor Area Extraction

The color properties in the peripheral area of a tumor are known to be of particular importance for classifying melanomas. Therefore, most studies on automated melanoma classification perform tumor area segmentation as an important preprocessing step. For this reason, many methods for automated tumor segmentation from dermoscopy images were proposed [9,12–15]. In this study, we applied our former method [9], the performance of which was evaluated with several state-of-the-art methods and confirmed [15]. Because of the space limitation, please refer to our original article for details.

3.2 Calculation of the Image Features

We designed and used a total of 120 color image features extracted from tumors in this study based on the consideration of the ABCD rule [3]. We named them "base-parameters." We emphasize here that we did not use image features that we could not easily understand, such as textures, frequency components, or coefficients calculated by known robust features, i.e., SIFT, SURF etc., in order to attach great importance to the readability of the "grounds" of the discrimination results. For calculating the base-parameters, we focused on the tumor body (T)and the periphery of the tumor (P), and the difference between the tumor and its surrounding skin (S-T) and that between the tumor and its periphery (T-P). In each part of the tumor, we calculated the mean (μ), minimum (min), maximum (max), standard deviation (σ), and skewness (skew) values for each RGB and HSV color channel, i.e., $5 \times 6 = 30$ features in each part. Note that the peripheral part of the tumor was defined as the region inside the tumor's border that has an area equal to 30 % of the tumor area as determined by a recursive dilation process applied to the outer border, working inward from the border of the extracted tumor [9].

3.3 Selection of the Important Features with RF

In this phase, we determined a set of important features for classifying melanomas, called "important-parameters," as a subset of the "base-parameters." In classification or machine learning problems, feature selection is a requisite process for

achieving robust and reliable systems. For example, the incremental stepwise method [16] has been widely used because of the theoretical evidence of its effectiveness; however, it does not always lead to an appropriate solution in the case of nonlinear systems, in particular those including iterative processes. Accordingly, we propose a new novel feature selection method that uses the mechanism of RF. First, we trained RF with the base-parameters and calculated the effectiveness of the parameter p, $\eta(k,p)$, in each decision tree. Then, we calculated the overall influence of the parameter p by using

$$\alpha(p) = \frac{\sum_k \eta(k,p)}{\sum_k \sum_p \eta(k,p)}. \tag{4}$$

It should be noted here that $\sum_p \alpha(p) = 1$ because of its normalized definition. In this study, we selected the number of decision trees and their depth to be 100 and 10, respectively. We selected features having $\alpha(p)$ equal to or larger than 0.01 as the important-parameters based on our preliminary experiments.

3.4 Training of RF

We trained the RF with the important-parameters. In the k-th node, given data are divided into two groups by means of the selected parameter $p^*(k)$ and associated threshold $\theta^*(k)$. The threshold $\theta^*(k)$ is determined by (1) generating several random values $\boldsymbol{\theta}(k)$ as the candidates of the threshold between the maximum and minimum value in terms of $p^*(k)$ of the incoming data, (2) calculating the amount of entropy reduction for each $\theta(k)$, and (3) selecting $\theta^*(k)$ with the largest entropy reduction. In addition, it is known that if the training dataset has a class bias, the performance of the built RF will be also biased. In this study, we introduced a reverse biased sampling and weighting technique to normalize the class bias in the training dataset [12].

3.5 Exhibition of the Classification Result and Decision Tree

Melanoma classification is performed using the built RF classifier. Note that the final decision was made by majority voting of each decision tree. We selected one decision tree with the highest average of $\alpha(p)$ among all trees yielding the majority result. The system provided the path from the root node to the leaf node, i.e., the final discrimination result, as the grounds for the result.

4 Experiment

In this study, we used a total of 1,148 dermoscopy images having an established diagnosis (168 melanomas and 980 nevi) from Naples, Graz, and Keio Universities. We obtained the important-parameters using the method described above. In order to confirm the effectiveness of our method, in particular its feature selection process, described in Sect. 3.3, we conducted three comparative experiments.

– RF trained with the base-parameters (performance baseline)
– RF trained with parameters selected using the stepwise method
– RF trained with the important-parameters (proposed method).

We used sensitivity and specificity as evaluation criteria and the classification performance of the RF was evaluated using a 10-fold cross-validation strategy.

5 Results

A total of 24 color features were selected as the important-parameters. Their top five features are summarized in Table 1. We determined that the number of decision trees and their depth were 50 and 8, respectively, according to the preliminary experiments. The final classification results for melanomas are summarized in Table 2. The proposed method (Experiment 3) attained the best performance values: SE = 79.8 % and SP = 80.7 %. Figure 1 shows an example of a test dermoscopy image (nevus) and its discrimination path in the built decision tree. Figure 2 shows its partial enlargement. In a rectangular (non-terminal) node in the figure, the upper line represents the number of data incoming to the node and the lower line represents the conditions for data division, i.e., inaction with the image feature $p^*(k)$ and the corresponding threshold $\theta^*(k)$). If the condition is satisfied, the algorithm follows the left link; otherwise, the right. The nodes having rounded corners in the lower part of figure represent the final discrimination result; gray and white indicate malignant (M) and benign (N) cases, respectively. The upper line indicates the number of data incoming to the node and the lower indicates those classified in each category. Note here that this number is compensated by the reciprocal ratio of malignant and benign

Table 1. Selected important-parameters (top 5)

Rank	Feature	Description	$\alpha(p)$ (%)
1	σ_G^P	Standard deviation of green in the periphery	2.3
2	μ_B^P	Average of blue in the periphery	2.2
3	σ_R^T	Standard deviation of red in the tumor	2.0
4	σ_V^{T-P}	Standard deviation of intensity between the tumor and its periphery	1.8
5	σ_G^T	Standard deviation of green in the tumor	1.6

Table 2. Comparison of melanoma discrimination performance

Experiments	Feature selection	No. of features	SE (%)	SP (%)
Experiment 1	no	120	69.9	70.5
Experiment 2	stepwise [16]	14	70.2	68.8
Experiment 3	RF (proposed)	24	79.8	80.7

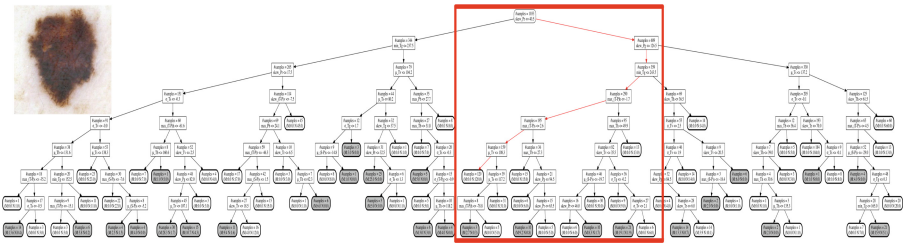

Fig. 1. Discrimination of a nevus with the built decision tree.

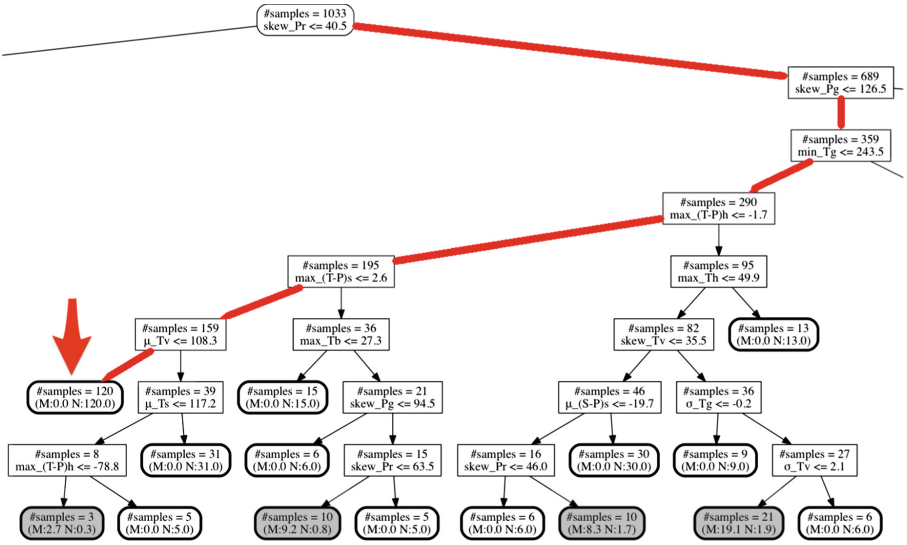

Fig. 2. Enlarged view of Fig. 1.

Table 3. Selected important-parameters (top seven)

Rank	Feature	Description	$\alpha(p)$ (%)
1	σ_V^T	Standard deviation of intensity in the tumor	21.9
2	$\max_B^P$	Maximum of blue in the periphery	11.5
3	skew_V^{T-P}	Difference of skew in intensity between the tumor and its periphery	10.7
4	$\min_G^T$	Minimum of green in the tumor	8.2
5	μ_B^T	Average of blue in the tumor	7.6
6	skew_V^T	Skew of intensity in the tumor	6.7
7	σ_G^T	Difference of standard deviation in green between the tumor and its periphery	4.2

cases, and each tree was trained with 90 % of the training data because of the 10 fold cross-validation strategy. The image features with larger importance $\alpha(p)$ in Experiment 3 are summarized in Table 3.

6 Discussion

We confirmed that the RF classifier with the proposed feature selection method (Experiment 3) attained the best classification performance of all the methods. When the incremental stepwise input selection was applied (Experiment 2), the classification performance was worse than that of Experiment 1. The incremental stepwise method is designed for selecting appropriate feature sets for single linear models with a statistical t-test. We considered this one of the best solutions for a single linear model, but the limited number of features largely reduced the degree of freedom of the parameter pools and accordingly caused the performance degradation of RF. RF requires a certain degree of freedom in a parameter pool; the proposed feature selection method provided this. In Table 3, it can be seen that the most significant feature was the standard deviation of the intensity in the tumor body (σ_V^T). We consider this represents the polychromism of the tumor caused by the variation in the depth of melanin from the epidermis. This is consistent with the diagnostic criteria used in clinical practice. The second and the following features represent the blue, intensity, or green (highly correlated with the intensity) property of the tumor or its periphery. The results are also considered to show that our proposed feature selection method selected image features that are important for classifying melanomas appropriately. In the RF algorithm, test data are inputted to all the built decision trees and in each tree inputted data items remain sorted by the condition described in each node until they reach the terminal. The proposed system shows the process of a typical tree as the grounds for the result. Again, because we attached the highest importance to building easy-to-read "grounds for the result," and to suppressing over-learning, we limited the number of layers of the tree, and the base-parameters comprise only color features. Because of these human-oriented limitations, on the one hand, the numerical classification performance of our system gave a slight advantage to other state-of-the-art methods that focus on only the classification performance, while on the other, its performance is comparable to that of expert dermatologists. However, our current system provides the grounds of the discrimination expressed as a tree with numerical values. We are currently investigating a method for converting the numerical value to a more intelligible linguistic expression. We will report the results of this endeavor and upload this function on our Internet melanoma classification system in the near feature. However, our method successfully attains a performance comparable to that of expert dermatologists.

7 Conclusion

In this paper, we proposed an automated melanoma discrimination system capable of offering both the discrimination results and their grounds by means of

utilizing a Random Forest (RF) technique. Owing to the introduction of our feature selection method for RF, our system is capable of providing not only a good classification performance, but also the grounds for the discrimination results in an intelligible format.Our system provides the grounds of discrimination in the tree expressions with numerical values. We are currently investigating to convert this into more intelligible linguistic expressions, and will incorporate this functionality into our Internet melanoma classification system in the near feature.

Acknowledgement. This research was partially supported by the Ministry of Education, Culture, Science and Technology of Japan (Grant in-Aid for Fundamental research program (C), 26461666, 2014-2017).

References

1. National Cancer Institute: SEER Cancer Statistics Review 1975–2012. NCI, Bethesda, MD (2015)
2. Soyer, H.P., Smolle, J., Kerl, H., Stettner, H.: Early diagnosis of malignant melanoma by surface microscopy. The Lancet **2**(8562), 803 (1987)
3. Stolz, W., Riemann, A., Cognetta, A.B., Pillet, L., Abmayr, W., Holzel, D., et al.: ABCD rule of dermoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. **4**(7), 521–527 (1993)
4. Argenziano, G., Fabbrocini, G., Carli, P., Giorgi, V.D., et al.: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch. Dermatol. **134**(12), 1563–1570 (1998)
5. Stolz, W., Falco, O.B., Bliek, P., Kandthaler, M., Burgdorf, W.H.C., Cognetta, A.B.: Color Atlas of Dermatoscopy, 2nd enlarged and completely revised edn. Blackwell Publishing, Berlin (2002)
6. Rubegni, P., Cecenini, G., Burroni, M., Perotti, R., Del'Eva, G., Sbano, P., et al.: Automated diagnosis of pigmented skin lesions. Int. J. Cancer **101**(6), 576–580 (2002)
7. Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., et al.: A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. **31**(6), 362–373 (2007)
8. Oka, H., Hashimoto, M., Iyatomi, H., Argenziano, G., Soyer, H.P., Tanaka, M.: Internet-based program for automatic discrimination of dermoscopic images between melanomas and Clark nevi. British J. Dermatol. **150**(5), 1041 (2004)
9. Iyatomi, H., Oka, H., Saito, M., Miyake, A., Kimoto, M., Yamagami, J., et al.: Quantitative assessment of tumor extraction from dermoscopy images and evaluation of computer-based extraction methods for automatic melanoma diagnostic system. Melanoma Res. **16**(2), 183–190 (2006)
10. Iyatomi, H., Oka, H., Celebi, M.E., Ogawa, K., Argenziano, G., Soyer, H.P., Koga, H., Saida, T., Ohara, K., Tanaka, M.: Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin. J. Invest. Dermatol. **128**(8), 2049–2054 (2008)
11. Iyatomi, H., Oka, H., Celebi, M.E., Tanaka, M., Ogawa, K.: Parameterization of dermoscopic findings for the Internet-based melanoma diagnostic system. In: Proceedings of the IEEE CIISP 2007, pp. 183–193 (2007)

12. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
13. Joel, G., Philippe, S.S., et al.: Validation of segmentation techniques for digital dermoscopy. Skin Res. Technol. **8**(4), 240–249 (2002)
14. Celebi, M.E., Iyatomi, H., Schaefer, G., Stoecker, W.V.: Lesion border detection in dermoscopy images. Comput. Med. Imaging Graph. **33**(2), 148–153 (2009)
15. Norton, K.A., Iyatomi, H., Celebi, M.E., Ishizaki, S., et al.: Three-phase general border detection method for dermoscopy images using non-uniform illumination correction. Skin Res. Technol. **18**(3), 290–300 (2012)
16. Hocking, R.R.: The analysis and selection of variables in linear regression. Biometrics **32**(1), 1–49 (1976)