Classifying Glaucoma with Image-Based Features from Fundus Photographs

Rüdiger Bock[1], Jörg Meier[1], Georg Michelson[2], László G. Nyúl[1], and Joachim Hornegger[1]

[1] Institute of Pattern Recognition, University of Erlangen-Nuremberg,
Martensstraße 3, 91058 Erlangen
{ruediger.bock,joerg.meier,laszlo.nyul}@informatik.uni-erlangen.de,
joachim.hornegger@informatik.uni-erlangen.de
[2] Department of Ophthalmology, University of Erlangen-Nuremberg,
Schwabachanlage 6, 91054 Erlangen
georg.michelson@augen.imed.uni-erlangen.de

Abstract. Glaucoma is one of the most common causes of blindness and it is becoming even more important considering the ageing society. Because healing of died retinal nerve fibers is not possible early detection and prevention is essential. Robust, automated mass-screening will help to extend the symptom-free life of affected patients. We devised a novel, automated, appearance based glaucoma classification system that does not depend on segmentation based measurements. Our purely data-driven approach is applicable in large-scale screening examinations. It applies a standard pattern recognition pipeline with a 2-stage classification step. Several types of image-based features were analyzed and are combined to capture glaucomatous structures. Certain disease independent variations such as illumination inhomogeneities, size differences, and vessel structures are eliminated in the preprocessing phase. The "vessel-free" images and intermediate results of the methods are novel representations of the data for the physicians that may provide new insight into and help to better understand glaucoma. Our system achieves 86 % success rate on a data set containing a mixture of 200 real images of healthy and glaucomatous eyes. The performance of the system is comparable to human medical experts in detecting glaucomatous retina fundus images.

1 Introduction

Glaucoma is one of the most common causes of blindness with a mean prevalence of 4.2 % for ages above 60 years. This disease is characterized by changes in the eyeground (fundus) in the region of the optic nerve head (ONH): (i) enlargement of the excavation, (ii) disc hemorrhage, (iii) thinning of the neuroretinal rim, (iv) asymmetry of the cup between left and right eye, (v) loss of retina nerve fibers, and (vi) appearance of parapapillary atrophy. It is induced by the progressive loss of retinal nerve fibers in the parapapillary region. Although those lost fibers cannot be revitalized and there is no possibility for healing glaucoma, the progression of the disease can be stopped [1].

Nowadays, diagnosis is commonly done by physicians who examine the eye fundus using an ophthalmoscope or digital retina images acquired by devices like the Heidelberg Retina Tomograph (HRT) [2] or the Kowa NonMyd fundus camera.

1.1   State of the Art

In the domain of retina image analysis, automated methods already exist for certain tasks, for instance determination of components of the eyeground (e.g. segmenting the vessels [3] or the ONH [4,5]). These measurements can be used for automated diagnosis of diseases such as diabetic retinopathy [6] or glaucoma. Existing image-based glaucoma detection methods work on HRT images. Swindale et al. [7,8] models a smooth two-dimensional surface that is fitted to the optic nerve head in topography images. Detection of glaucomatous damages can be done via global shape measures of the optic disc (cup and disc area, height variation in HRT images) [9]. This global shape approach is compared with a sector-based analysis by Iester et al. [10]. Zangwill evaluated optic disc parameters and additional parapapillary parameters via Support Vector Machine (SVM) for detecting glaucoma [11]. Greaney states that the detection of glaucoma via separately applied, shape-based methods on different modalities (confocal laser scanning ophthalmoscopy, scanning laser polarimetry, optical coherence tomography) is not better than qualitative assessment of the optic disc by ophthalmologists [12]. All of these shape approaches assume a valid segmentation of the optic disc. However, segmentation based techniques have one major drawback: small errors in segmentation may lead to significant change in the measurements and thus the estimation and diagnosis.

1.2   Our Approach

We build a robust, automated glaucoma detection system using color fundus images in a data-driven way. Therefore, image-based features are provided that are new in the domain of glaucoma detection. This, so called appearance based, approach is well-known from object and face recognition [13,14]. The technique is based on statistical evaluation of the data and does not depend on explicit outlining of the optic disc, as required for global or sector-based shape analysis. Consequently, preprocessing and image-based feature extraction has a major influence on the classification process.

This work shows the influence of different image-based features on the accuracy of glaucoma classification from fundus images. We analyze different types of features (pixel intensity values, textures, spectral features, and parameters of a histogram model) that are intended to capture glaucomatous structures and evaluate the results using three different classifiers (naive Bayes classifier, k-nearest neighbor, and Support Vector Machine). They were used to classify the computed features as is, in combination with an attribute preselection method, and with an iterative attribute selection by AdaBoosting. The combination of features is also considered.

2 Methods

The image processing is structured in a standard 3-stage pipeline: (i) preprocessing, (ii) image-based feature extraction, (iii) classification. In this work, the preprocessing steps are not changed during the experiments.

2.1 Preprocessing

Our previous studies [15] showed that appearance-based approaches perform better on images with less disease independent variations. For this reason, a normalization of the three major variations is applied.

The inhomogeneous illumination caused by deviations in the complex acquisition process are compensated by robust homomorphic surface fitting [6]. Also, the vessel branches in the images vary much in size, location and shape among individual cases and introduce a high variance in the data that suppresses variations due to the disease itself. Thus, we roughly segment blood vessels and spatially inpaint them to gain a "vessel-free" image. These images, as shown in Fig. 1, provide a novel image representation with irrelevant parts excluded that support physicians in diagnosing glaucoma.

Fig. 1. Vessel inpainting on color fundus image: Original color fundus image (left), image with vessel mask overlayed in black (center), "vessel-free" fundus image (right)

The neuroretinal rim is the most important region for detecting glaucoma [16]. We normalize the images such that the ONH is centered in all images and appears in the same size. The images are scaled to an uniform size of 128×128 pixels.

2.2 Image-Based Feature Extraction

We propose four feature extraction methods that provide complementary information with different spatial and frequency resolutions. Their influence on glaucoma classification is evaluated in this paper.

Pixel Intensity Values: The standard appearance based approach takes pixel intensity values directly as a high dimensional feature vector [13] as input of a dimension reduction algorithm. We considered principal component analysis (PCA) as an unsupervised and linear discriminant analysis (LDA) as a supervised method to reduce dimensionality. As evaluated in [15], thirty principal components capture at least 95 % of data variation.

Textures: The local spatial and spatial-frequency information is characterized by textures. To capture the structural changes caused by glaucoma, we provide a set of Gabor filter banks [17] on preprocessed images. The filter is performed for rotation angles $\theta = 0°, 45°, 90°, 135°$ and frequencies $u_0 = 2^k \sqrt{2}$ for $k = 0, \ldots, 6$. The dimension of each of the computed 28 filter responses ($128 \times 128\,\mathrm{px}$) is reduced by PCA separately. The resulting eigenspaces are then concatenated to a single $(28 \cdot 30)$-dimensional space.

FFT Coefficients: The frequency spectrum contains global frequency information that is translation invariant. We calculate the real and imaginary response as well as the magnitude of the coefficients of the Fast Fourier Transform (FFT). A dimension reduction via PCA is performed on the three responses and combined to a $(3 \cdot 30)$-dimensional space.

Histogram Model: Histograms provide a compact summary of the data distribution in an image. The histogram of an image is relatively invariant to translation and rotation of objects. Comparing histograms of different images is particularly well applicable to the problem of recognizing global intensity changes. The histograms of the preprocessed retina images show three major structural parts corresponding to the background, the papilla rim, and the cup.

The expected variation in the images because of the disease is also represented in the histograms. The increasing cup area and the decreasing rim area cause a shift of the intensity distribution towards higher values. We fit a Gaussian mixture model of three normal distributions to the histogram by a maximum likelihood estimation and the computed distribution parameters, namely the mean, variance and weight, serve as features. The $10\,\%$ and $90\,\%$ quantiles and the maximum of the histogram are also taken into consideration.

2.3 Classification

Classifiers: The ability of each image-based feature extraction method to separate glaucoma and non-glaucoma cases is quantified by the results of three classifiers. Classifiers achieve good results if their underlying separation model fits well to the distribution of the sample data. As the underlying data distribution is unknown, we tested different classifiers.

Naive Bayes Classifier: This probabilistic classifier directly applies the Bayes rule to determine the probability of a test sample belonging to a class. Three assumptions are made: the feature data is normally distributed, the predictive attributes are conditionally independent given the class, and no hidden or latent attributes influence the prediction process [18].

k-Nearest Neighbor Classifier (k-NN): The k-NN classifier as instance-based classifier, does not assume a specific distribution of the feature data. It adapts well to the sample data, but also tends to overfit. It is also sensitive to noise and to irrelevant features. It is applied with $k = 5$ neighbors.

Support Vector Machine: This linear classifier determines a maximum-margin and soft hyperplane that best separates the considered classes. The data is normalized and transformed via the non-linear radial basis kernel. We use the ν-SVM with penalization parameter $\nu = 0.5$ and cost-parameter $c = 1$ [19].

Classification Enhancement: The feature data distribution might not optimally fit to the classifiers' data model. We analyze the effect of two known methods to improve classification result.

(i) Feature selection removes attributes from the initial set. Features that are highly correlated to a class and have a low correlation to other attributes are kept [20] and the reduced set is used for classification.

(ii) AdaBoosting is a classification scheme to improve classification results. The method iteratively applies one arbitrary classifier. AdaBoosting is able to improve results especially of weak learners on real-world data and is robust to overfitting [21].

Feature Combination: To further improve the classification correctness and robustness, we investigated two ways of combining the four image-based feature types.

(i) Feature Merging concatenates all available feature spaces to a new high dimensional space (970 dim.) that is used for classification.

(ii) 2-stage classification applies the probability score of belonging to the glaucoma class, obtained from each of the four classifiers, as new feature vector input to another classifier.

3   Evaluation

For evaluation, we took images from the Erlangen Glaucoma Registry (EGR) that contains thousands of records of multi-modal fundus images from a long-term screening study. Diagnosis was made by an ophthalmologist based on anamnesis, image data and other measurements. The images were acquired by a Kowa NonMyd α-D digital fundus camera that produced lossless compressed RGB photographs of size 1600×1216 pixels, using a 20° field of view and nasal positioning (papilla-centered).

We evaluated the above described image-based feature extraction methods and classifiers on a test set of 100 preprocessed images (50 healthy and 50 glaucomatous). With this set, the measures were calculated for cross-validation tests and for classification experiments with separated training and test data. The PCA/LDA models, i.e. the eigenimages, as well as the training of the classifiers in the case of separated training and test sets was done with another image set of 100 images (also 50 healthy and 50 glaucomatous cases).

For all experiments we computed the overall classification correctness and the F-measure for healthy (F_h) and glaucomatous eyes (F_g), which is the harmonic mean of sensitivity and precision. To mark promising and robust configurations,

Table 1. Classification performance of the four feature extraction methods. Configurations with "best"-criterion are labeled bold.

Data	Classifier	Structure	Cross-validation			Train-Test		
			Correct (%)	F_h	F_g	Correct (%)	F_h	F_g
PCA on intensities (30 dim.)	Bayes	nothing	73	0.72	0.74	48	0.43	0.52
		AdaBoost	75	0.76	0.74	64	0.65	0.63
		FeatureSel	77	0.75	0.79	63	0.67	0.58
	kNN	nothing	77	0.75	0.79	68	0.75	0.57
		AdaBoost	74	0.75	0.73	70	0.75	0.63
		FeatureSel	82	0.82	0.82	70	0.75	0.62
	SVM	**nothing**	**83**	**0.81**	**0.85**	**81**	**0.83**	**0.78**
		AdaBoost	80	0.78	0.82	79	0.82	0.75
		FeatureSel	85	0.85	0.85	73	0.76	0.69
PCA on textures (840 dim.)	Bayes	nothing	69	0.67	0.70	44	0.44	0.44
		AdaBoost	69	0.70	0.67	70	0.67	0.72
		FeatureSel	76	0.76	0.76	60	0.62	0.57
	kNN	nothing	55	0.21	0.69	64	0.73	0.45
		AdaBoost	64	0.49	0.72	66	0.74	0.51
		FeatureSel	80	0.76	0.83	73	0.77	0.67
	SVM	nothing	67	0.57	0.73	60	0.71	0.35
		AdaBoost	60	0.41	0.70	76	0.77	0.74
		FeatureSel	**80**	**0.78**	**0.82**	**81**	**0.83**	**0.79**
PCA on FFT (90 dim.)	Bayes	nothing	74	0.73	0.75	47	0.40	0.52
		AdaBoost	83	0.83	0.83	59	0.55	0.62
		FeatureSel	78	0.77	0.79	72	0.76	0.66
	kNN	nothing	75	0.71	0.78	66	0.74	0.50
		AdaBoost	74	0.74	0.74	71	0.77	0.60
		FeatureSel	83	0.82	0.84	69	0.74	0.62
	SVM	**nothing**	**76**	**0.74**	**0.77**	**76**	**0.79**	**0.71**
		AdaBoost	77	0.76	0.78	72	0.76	0.67
		FeatureSel	83	0.82	0.84	73	0.77	0.67
Histogram model (10 dim.)	Bayes	nothing	71	0.69	0.73	58	0.60	0.55
		AdaBoost	73	0.71	0.75	65	0.69	0.60
		FeatureSel	71	0.69	0.73	45	0.32	0.54
	kNN	nothing	81	0.80	0.82	54	0.50	0.57
		AdaBoost	72	0.67	0.77	54	0.50	0.57
		FeatureSel	69	0.69	0.68	42	0.39	0.44
	SVM	**nothing**	**73**	**0.72**	**0.74**	**61**	**0.61**	**0.60**
		AdaBoost	80	0.80	0.80	85	0.00	0.92
		FeatureSel	70	0.69	0.71	39	0.30	0.46
LDA on intensities (30 dim.)	Bayes	nothing	78	0.78	0.78	55	0.52	0.58
		AdaBoost	77	0.78	0.76	68	0.71	0.65
		FeatureSel	79	0.78	0.80	67	0.71	0.61
	kNN	nothing	79	0.78	0.80	68	0.75	0.57
		AdaBoost	81	0.80	0.83	74	0.78	0.68
		FeatureSel	84	0.84	0.84	69	0.75	0.60
	SVM	**nothing**	**82**	**0.80**	**0.84**	**76**	**0.80**	**0.71**
		AdaBoost	78	0.78	0.78	68	0.75	0.57
		FeatureSel	80	0.79	0.81	69	0.75	0.60

we defined a "best"-criterion for each feature extraction set. The best configuration within a set has a $F_g \geq 0.60$ and $F_h \geq 0.60$ in the cross-validation test as well as in the separated training and test sets experiments and maximum sum of F-measures $F_g + F_h$.
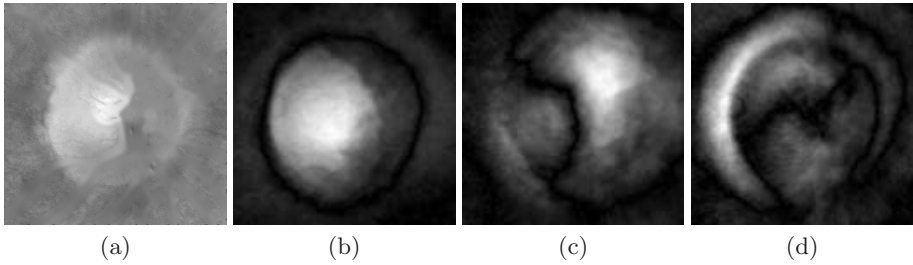
| (a) | (b) | (c) | (d) |

Fig. 2. (a) "Vessel-free" fundus image, (b)-(c) the first three absolute eigenimages of PCA on intensities. Bright regions indicate high influence on the features.

Table 2. Classification performance of merged features and 2-stage classification

Data	Classifier	Structure	Cross-validation			Train-Test		
			Correct (%)	F_h	F_g	Correct (%)	F_h	F_g
Feature Merging (970 dim.)	Bayes	nothing	71	0.69	0.72	45	0.43	0.47
		AdaBoost	79	0.80	0.78	63	0.58	0.67
		FeatureSel	76	0.76	0.76	68	0.71	0.64
	kNN	nothing	60	0.33	0.71	63	0.72	0.45
		AdaBoost	71	0.69	0.73	66	0.74	0.51
		FeatureSel	81	0.79	0.83	67	0.74	0.56
	SVM	nothing	72	0.65	0.77	61	0.72	0.38
		AdaBoost	61	0.42	0.71	62	0.72	0.71
		FeatureSel	**84**	**0.83**	**0.85**	**80**	**0.82**	**0.77**
2-stage Classification (4 dim.)	Bayes	nothing	84	0.85	0.86	80	0.82	0.76
		AdaBoost	82	0.82	0.82	80	0.82	0.77
	kNN	nothing	81	0.80	0.82	80	0.82	0.77
		AdaBoost	80	0.80	0.80	78	0.81	0.74
	SVM	nothing	85	0.86	0.84	80	0.82	0.77
		AdaBoost	**86**	**0.83**	**0.88**	**80**	**0.82**	**0.77**

In the first step, we tested the four feature extractions with the different classifier types as described in Section 2. Each feature configuration was applied as is, in combination with attribute selection or AdaBoosting. The classification results are given in Table 1. We also computed features based on a LDA model (trained with the separated training set) and classified them according to our scheme. The result is shown in Table 1. The absolute eigenimages generated by PCA on intensities (see Fig. 2) show regions of the fundus with high influence to the features. Those regions might point to relevant glaucomatous areas and help in understanding glaucoma.

In the second step, we evaluated the performance of the feature combinations. Feature merging results are stated in the first block of Table 2. Two stage classification combines the four best classifiers for each feature extraction method. For each feature extraction the class probability for glaucoma was taken as an input to a second classification step. As these features are only 4 dimensional, there was no need for feature selection. The classification result is shown in the second block of Table 2.

4   Results and Discussion

The classification performance using each feature extraction method separately shows that the correctness varies between 55 % and 85 % in cross-validation. Also each feature extraction method itself has varying classification correctness and F-measures for the different classifier configurations. The SVM separates the features most robustly and is always part of configurations labeled with the "best"-criterion. The configurations with "best"-criterion achieve F-measures between 0.72 and 0.81 for healthy case and between 0.74 and 0.85 for glaucomatous case in case of cross-validation. They are always using SVM for classification. Only in case of high dimensional feature space as with PCA on textures (840 dim.) feature selection is necessary to avoid problems with the curse of dimensionality and to achieve similar success rates as PCA on intensities. Although the PCA on pixel values ($F_g = 0.85$) and on texture ($F_g = 0.82$) shows slight better results than PCA on FFT ($F_g = 0.77$) and the histogram model ($F_g = 0.74$), all configurations show a reasonable discriminative power. Comparing the two dimension reduction techniques, LDA shows a smaller variance in the results than PCA. The SVM also classifies the LDA features best.

In case of the feature merging, the highest success rate and F-measures ($F_g = 0.85$, $F_h = 0.83$) are obtained if a feature selection is done before using the SVM in case of cross-validation. In 2-stage classification, the class-probabilities of the "best"-labeled classifier configurations are used as second stage features. This scheme shows success rates with F-measures over 0.80 for all classifier configurations in case of cross-validation. Classification on separate training and test set shows consistent, but slight inferior F-measures. The highest success rate of all experiments (86 %) is gained by SVM with AdaBoosting with $F_h = 0.83$ and $F_g = 0.88$.

As stated in [12], experienced observers achieve an average $F_g = 0.79$ and $F_h = 0.91$ by qualitative assessment of optic nerve head stereophotographs (63 normal and 29 glaucomatous subjects). Regarding classification on separate test and training set, we gain a slightly inferior performance ($F_g = 0.77$) while we get $F_h = 0.82$ for normals.

5   Conclusion

We presented a novel automated glaucoma classification system using digital fundus images. In contrast to the commonly used segmentation based measurements, it is purely data-driven and uses image-based features that are new in the domain of glaucoma recognition. We evaluated several combinations of image-based features and classifier schemes on a set of 200 real fundus images. The 2-stage classification with SVM produced 86 % success rate. The performance of the fully automatic system presented here is comparable to medical experts in detecting glaucomatous eyes and it could be used in mass-screenings. The important features automatically identified by the methods also provide a novel representation of the data for the physicians and may help to better understand glaucoma.

Acknowledgments

References

1. Sivalingam, E.: Glaucoma: An overview. J. Ophthalmic. Nurs. Tech. 15(1), 15–18 (1996)
2. Malinovsky, V.E.: An overview of the Heidelberg Retina Tomograph. J. Am. Optom. Assoc. 67(8), 457–467 (1996)
3. Staal, J., Abràmoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imag. 23(4), 501–509 (2004)
4. Hoover, A., Goldbaum, M.: Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE Trans. Med. Imag. 22(8), 951–958 (2003)
5. Chrástek, R., Wolf, M., Donath, K., Niemann, H., Paulus, D., Hothorn, T., Lausen, B., Lämmer, R., Mardin, C., Michelson, G.: Automated segmentation of the optic nerve head for diagnosis of glaucoma. Med. Image Anal. 9(4), 297–314 (2005)
6. Narasimha-Iyer, H., Can, A., Roysam, B., Stewart, C.V., Tanenbaum, H.L., Majerovics, A., Singh, H.: Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy. IEEE Trans. Biomed. Eng. 53(6), 1084–1098 (2006)
7. Swindale, N.V., Stjepanovic, G., Chin, A., Mikelberg, F.S.: Automated analysis of normal and glaucomatous optic nerve head topography images. Invest. Ophthalmol. Vis. Sci. 41(7), 1730–1742 (2000)
8. Adler, W., Hothorn, T., Lausen, B.: Simulation based analysis of automated, classification of medical images. Methods Inf. Med. 43(2), 150–155 (2004)
9. Uchida, H., Brigatti, L., Caprioli, J.: Detection of structural damage from glaucoma with confocal laser image analysis. Invest. Ophthalmol. Vis. Sci. 37(12), 2393–2401 (1996)
10. Iester, M., Swindale, N.V., Mikelberg, F.S.: Sector-based analysis of optic nerve head shape parameters and visual field indices in healthy and glaucomatous eyes. J. Glaucoma. 6(6), 370–376 (1997)
11. Zangwill, L.M., Chan, K., Bowd, C., Hao, J., Lee, T.W., Weinreb, R.N., Sejnowski, T.J., Goldbaum, M.H.: Heidelberg retina tomograph measurements of the optic disc and parapapillary retina for detecting glaucoma analyzed by machine learning classifiers. Invest. Ophthalmol. Vis. Sci. 45(9), 3144–3151 (2004)
12. Greaney, M.J., Hoffman, D.C., Garway-Heath, D.F., Nakla, M., Coleman, A.L., Caprioli, J.: Comparison of optic nerve imaging methods to distinguish normal eyes from those with glaucoma. Invest. Ophthalmol. Vis. Sci. 43(1), 140–145 (2002)
13. Hornegger, J., Niemann, H., Risack, R.: Appearance-based object recognition using optimal feature transforms. Pattern Recogn 2(33), 209–224 (2000)

14. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Comput. Surv. 35(4), 399–458 (2003)
15. Meier, J., Bock, R., Michelson, G., Nyúl, L.G., Hornegger, J.: Effects of preprocessing eye fundus images on appearance based glaucoma classification. In: Procceedings of International Conference on Computer Analysis of Images and Patterns (2007) (accepted for publication)
16. Lester, M., Garway-Heath, D., Lemij, H.: Optic Nerve Head and Retinal Nerve Fibre Analysis. European Glaucoma Society (2005)
17. Jain, A., Farrokhnia, F.: Unsupervised texture segmentation using Gabor filters. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 14–19. IEEE Computer Society Press, Los Alamitos (1990)
18. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artifical Intelligence, San Mateo, pp. 338–345. Morgan Kaufmann Publishers, San Francisco (1995)
19. Chen, P.H., Lin, C.J., Schölkopf, B.: A tutorial on ν-support vector machines. Applied Stochastic Models in Business and Industry 21(2), 111–136 (2005)
20. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand (1999)
21. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)