Pamukkale Univ Muh Bilim Derg, 25(5), 596-602, 2019

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Continuous time threshold selection for binary classification on polarized data

Kutuplaştırılmış veri üzerinde ikili sınıflandırma için sürekli zamanlı eşik değeri belirleme

Ali SAGLAM[1*], *Nurdan AKHAN BAYKAN*[2]

[1,2]Deparment of Computer Engineering, Engineering and Natural Science Faculty, Konya Technical University, Konya, Turkey.
asaglam@ktun.edu.tr, nbaykan@ktun.edu.tr

Received/Geliş Tarihi: 07.08.2018, Accepted/Kabul Tarihi: 22.11.2018
doi: 10.5505/pajes.2018.26125
* Corresponding author/Yazışılan Yazar
Research Article/Araştırma Makalesi

Abstract

Binary classification is used to distinguish some of the data elements from others in a meaningful way according to certain characteristics. Supervised classification techniques often use the ground-truth data, which assists to determine the distinctive characteristics of the elements to be extracted from the data. These techniques also generate new features for all of the data using the current features in accordance with the ground-truth data. One of the purposes of generating new features is to polarize the data elements (to be extracted and others) toward the separate pools on a coordinate axis for binary classification. In this way, the binary classification process is easy using only a threshold value on the axis. In this work, the Linear Discriminant Analysis (LDA) is used to polarize the data and a threshold selection algorithm is proposed, which use the harmonic mean F-score values of the binary classification outputs resulting from some specific threshold values. The key condition in the proposed method is that the most suitable threshold must give the best classification score (F-score value) and other threshold values must give lower classification scores as they become distant from the best threshold value (move away toward the ends of the axis). The proposed method is experimented for binary classifications of some meaningful elements on a remote sensing image taken from a 2D semantic labelling dataset that has the ground-truth images. The proposed method convergences the best threshold value continuously in logarithmic time.

Keywords: Classification, Threshold, LDA, F-score

Öz

İkili sınıflandırma, veri elemanlarından bir kısmını belirli karakteristiklerine göre diğerlerinden anlamlı bir şekilde ayırmak için kullanılmaktadır. Denetimli sınıflandırma teknikleri ise genellikle veriden çıkarılacak elemanların ayırt edici karakteristiklerini belirlemeye yardımcı olan referans veriyi kullanmaktadır. Bu teknikler aynı zamanda mevcut özellikleri kullanarak bütün veri için referans veriye uygun olarak yeni özellikler oluşturmaktadır. Yeni özellikler oluşturmanın amaçlarından birisi de çıkarılacak veri elemanlarını ve diğerlerini ikili sınıflandırma için bir koordinat ekseni üzerinde ayrı kutuplara doğru kutuplaştırmaktır. Bu şekilde, sadece bir eksen üzerinde eşik değeri kullanarak, ikili sınıflandırma işlemi kolaylaşmaktadır. Bu çalışmada, veriyi kutuplaştırmak için doğrusal ayrıştırma analizi (DAA) kullanılmış ve bazı belirli eşik değerleriyle elde edilen ikili sınıflandırma çıktılarının harmonik ortalama F-score değerlerini kullanan bir eşik değeri belirleme algoritması önerilmiştir. Önerilen metottaki anahtar durum, en uygun eşik değeri en iyi sınıflandırma başarısını (F-score değerini) vermeli ve diğer eşik değerleri en iyi eşik değerinden uzaklaştıkça (eksenin iki ucuna doğru ilerledikçe) daha düşük sınıflandırma başarısını vermelidir. Önerilen metot, referans görüntüleri de içeren bir 2D anlamsal etiketleme veri kümesinden alınan bir uzaktan algılama görüntüsü üzerinde bazı anlamlı verilerin ikili sınıflandırması için uygulanmıştır. Önerilen metot en iyi eşik değerine sürekli zamanlı olarak belirlenen örnekleme sayısına ve sonlanma ölçütüne göre logaritmik zamanda yakınsamaktadır.

Anahtar kelimeler: Sınıflandırma, Eşik değeri, DAA, F-score

1 Introduction

In binary classification process, the data elements are grouped into two classes [1],[2] as the background and the foreground [3]. The foreground consists of the intended data elements to be extracted from the data and the background consists of other data elements outside the foreground. It is presumed that the foreground elements have some distinctive features that separate them from the background elements. These features may be the attribute values of the data elements, the pattern properties between the elements and the shape characteristics formed by a combination of some of the elements. In this study, we use only individual attribute values of the elements for binary classification.

Linear Discriminant Analysis (LDA) is a supervised method which makes easy to discriminate the intended elements from the others in the data [4],[5]. LDA uses the attributes of the data and assigns a weight value for each attribute; so that the data elements which belong to two different classes are polarized to two separate ends of the coordinate axis that occurs through the new values derived from the sum of the new weighted attribute values. In this way, LDA reduces the number of attributes of each element to one attribute [6]. For this purpose, LDA needs the training data and its ground-truth to define the distinctive features of the elements to be separated. The threshold selection is an endeavor subject for the binary classification of multi-dimensional data in the literature [7],[8]. Depending on the new polarized one-dimensional attribute values, the threshold value can also be determined easily using the training data and its ground-truth. After the weight values and the threshold are obtained, these values are applied on the test data without using any ground-truth data to obtain the new distinctive feature values [3].

The polarized data elements allow the use of one threshold value in different ways. Saglam and Baykan used the F-score values of the binary segmentation outputs resulting from the application of every integer in a range as threshold on the LDA values which are previously normalized in the range in their

study [3]. They firstly normalized the LDA values in the range 0-255 and discrete the values to integers. After computing the F-score values of all 256 threshold values, the best threshold value which gives the maximum F-score values is selected as the best threshold value. In this study, we converge the best threshold value continuously by trying fewer and non-discrete values as different from the reference study [3].

The proposed method, which allows achieving the best threshold, depends on the distribution of the measurement values of the classification successes. As the difference between the used threshold and the best threshold increases, the classification success of the used threshold must decrease. We examine some correctness measurements for classification; those are Accuracy and F-score [9],[10]. Accuracy is the rate of the number of data elements detected as true classes to all the data elements. F-score is the harmonic mean of the precision and the recall values [11]. F-score value indicates the success of one class. We call this detected region as foreground, while the remaining region as background. Precision is the rate of the true detected foreground region to the entire detected region. Recall is the rate of the true detected foreground region to the objective foreground region. While evaluating the binary classification, it is intended that the two classes (foreground and background) must reach the optimum F-score values. In literature, the optimum F-score values for multi-classes (including binary classification) are widely thought as the F-score values that give the maximum mean F-score value. In this study, we also examined the maximum harmonic mean F-score value and used this measurement as classification correctness, because this measurement provide the desired condition (a hill-climbing F-score distribution) for the proposed method. The harmonic mean F-score decreases while the threshold used moves away from the best threshold that gives the highest harmonic mean F-score.

We used a high resolution (1996 × 1995 pixels) remote sensing image to examine the classification correctness measurements and to apply the proposed threshold selection method. Finally, we checked that the result of the proposed method roughly matches the result of the discrete time method used in the reference study. One of the two advantages of the proposed method to the discretization approach is to be faster convergence to the best threshold value. The other advantage is allowing the determination of the convergence degree without relying on any discrete values such as integers.

2 Related works

In the past years, the threshold selection problems are dealt many times to solve different machine learning algorithms especially binary and multi-class classification problems. On the other hand, many performance measurement methods have been also used in many threshold selection algorithms.

Baldi et al. derived a few learning-based threshold selection algorithms by optimizing the correlation coefficient [12]. In their study, they used the Accuracy measurement and the relation between the reference data and the predictions of the probability between 0 and 1 using Hamming and Euclidean distances, which reflect the confidence degree of the predictions. The algorithms selected the threshold in the range 0-1. They firstly used the algorithm for binary classification; and then, they adapted the algorithm for multi-class problems.

Freeman and Moisen use a threshold optimizing method, instead of using the traditional value 0.5 in the range 0-1, for the classification of 13 species on a mountain [13]. They optimized 11 thresholds optimization criteria using the Accuracy of the surface model.

Sokolova and Lapalme analyzed 24 performance measurements for binary and multi-class classification according to their invariance and non-invariance properties [11]. They showed that the results of performance measurements can depend on the invariance properties of the measures. They also analyzed the applicability of performance measures on different text classification. They found out that the classification requires different performance measures depending on the text type to be classified.

Lipton et al. demonstrated the theoretical and empirical results that show the features of the F-score measure for binary and multi-class classification [7]. They used the best F-score value to specify the threshold for classification.

Sanchez considered a binary classifier as if the classifier is a player in a zero-sum game. [8]. The minimax principle from game theory was used to specify the optimal threshold by maximizing the Accuracy of the probabilities.

In this study, a threshold selection algorithm that runs on polarized data and uses the harmonic mean F-score values of two classes is proposed for binary classification. The binary classification can also be adapted for multi-class problems as seen in [3].

3 The binary classification with LDA

LDA transfers the data onto a one-dimensional coordinate axis by polarizing the data on the two ends of the axis. With the polarizing process, the distinguishable feature of the data for two classes becomes highest [3].

LDA firstly computes the sum of the intra-covariance matrices of two classes ($\boldsymbol{Cov_{intra}}$) as seen in Eq. 1.

$$\boldsymbol{Cov_{intra}} = \sum_{x_i \in fg} (\boldsymbol{x}_i - \boldsymbol{\mu}_{fg})(\boldsymbol{x}_i - \boldsymbol{\mu}_{fg})^T + \sum_{x_j \in bg} (\boldsymbol{x}_j - \boldsymbol{\mu}_{bg})(\boldsymbol{x}_j - \boldsymbol{\mu}_{bg})^T \quad (1)$$

In Eq. 1, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the attribute vectors of the elements of the foreground (fg) and background (bg) classes respectively. $\boldsymbol{\mu}_{fg}$ refers the mean attribute vectors of the foreground, while $\boldsymbol{\mu}_{bg}$ refers the mean attribute vectors of the background.

After computing $\boldsymbol{Cov_{intra}}$, polarizing the two classes requires only the calculation in Eq. 2. LDA values are calculated by multiplying the attribute value vectors by the vector **w** as seen in Eq. 3.

$$\boldsymbol{w} = \boldsymbol{Cov_{intra}^{-1}}(\boldsymbol{\mu}_{fg} - \boldsymbol{\mu}_{bg}) \quad (2)$$

$$LDA_n = \boldsymbol{w}.\boldsymbol{x}_n \quad | \quad \forall \boldsymbol{x}_n \in Data \quad (3)$$

In Eq. 2, the order of $\boldsymbol{\mu_{fg}}$ and $\boldsymbol{\mu}_{bg}$ is important. In the order in Eq. 2, the higher LDA values than the threshold value are considered as the elements of the foreground as in Eq. 4. If it is the reverse, the higher LDA values than the threshold value are considered as the elements of the background.

$$\begin{aligned} \text{foreground} &\Leftarrow LDA_n \geq threshold \\ \text{background} &\Leftarrow LDA_n < threshold \end{aligned} \quad (4)$$

4 Measurement of the classification accuracy

In the literature, the measurement of the classification success is usually ensured with Accuracy and F-score values. To calculating these values, True-Positive (TP), True-Negative (TN), False-Positive (FP) and False-Negative (FN) values must be calculated as defined in the sources [8],[11]. The regular representation of these definitions is presented in Table 1.

Table 1: The regular representation of the classification regions which the elements belong to.

		Detected Class	
		YES	NO
True Class	YES	TP	FN
	NO	FP	TN

4.1 Accuracy

Accuracy measurement is one of the best-known classification accuracy measurement methods. Due to the fact that accuracy focuses only true detected region, it is easy to implementation to multi-classification measurement. The general formulation of the accuracy is seen as Eq. 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

If we refer the true detected elements of foreground as TP_1 and the true detected elements of background as TP_2, the implementation of Accuracy to binary classification can be calculated as Eq. 6.

$$Accuracy = \frac{TP_1 + TP_2}{All\ the\ elements} \quad (6)$$

4.2 Arithmetic mean F-score

F-score measurement is another measurement method which is widely used in the classification field. It considers the false detected elements besides the true detected elements. F-score (Eq. 7) is the harmonic mean of Precision (Eq. 8) and Recall (Eq. 9) values.

$$F - score = 2.\frac{Precision.Recall}{Precision + Recall} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

F-score measures the success of the classification of one class. In multi-class classification, the F-score value is calculated for each class. To measure the overall classification, the arithmetic mean of the F-scores is commonly used in the literature [9],[10]. For the binary classification, the arithmetic mean of the F-scores can be calculated in Eq. 10, where F_1 refers the F-score of foreground and F_2 refers the F-score of background.

$$Arithmetic\ Mean\ F - score = \frac{F_1 + F_2}{2} \quad (10)$$

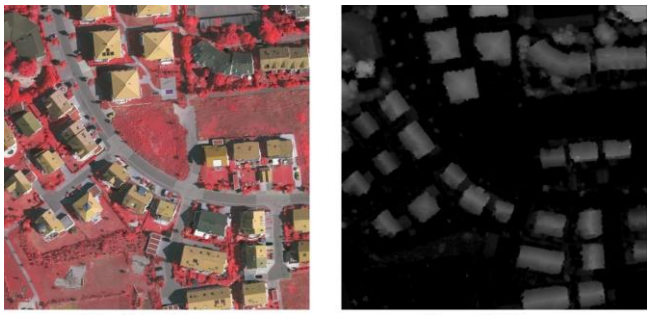
4.3 Harmonic mean F-score

Another overall F-score measurement approach used in this study is the harmonic mean of the foreground F-score (F_1) and the background F-score (F_2) and calculated as in Eq. 11.

$$Harmonic\ Mean\ F - score = 2.\frac{F_1.F_2}{F_1 + F_2} \quad (11)$$

5 The data used and the comparisons of the measurement methods

The data used for testing in this study is a high resolution (1996 × 1995 pixels) remote sensing image obtained from a 2D semantic labeling dataset [14],[15]. The dataset was captured over Vaihingen in Germany and carried out by the German Association of Photogrammetry and Remote Sensing (DGPF) using an Intergraph / ZI DMC used for Digital Aerial Images [16]. Each pixel in the image has the attributes IR (infrared), red (R), green (G), digital surface maps (DSM) and normalized DSM (nDSM) (Figure 1). The image includes the pixels that belong to some classes such as "Road", "Building", "Vegetation" and "Tree". In this work, we classify this image for binary classification; for example, the pixel belongs to "Road" or not.



The IR-R-G image | The nDSM image

Figure 1: The image used in this study.

We first calculated the LDA values of pixels for each class separately according to two classes (foreground and background). After that, we normalized the LDA values for each class to the range 0-255 and discretized them to integers as in the former studies. Finally, we computed the classification success of each integer in the range 0-255 as a threshold to classify the pixels to two classes and selected the threshold that gives the highest measurement value as the best threshold value. This threshold selection method [3] can be seen as a discrete threshold selection method.

In Table 2, the best threshold values are shown for the different classes and measurement methods. Looking the table, the threshold values are nearly the same except the class "Tree". In Figure 2, the comparison of the binary segmentation classification of the class "Tree" is seen visually. Looking to the Figure 2, it is seen that the Accuracy method gives under classification, whereas the classification according to Arithmetic and Harmonic mean F-score methods give over classification for the foreground class of "Tree". In this case, it is difficult to comment on which one is superior. For this reason, it can be ignored the difference between the best threshold values for measurement selection.

Table 2: The best discrete threshold values of 0 to 255.

	Threshold Values			
Measurement Method	Road	Build.	Veget.	Tree
Accuracy	164	119	189	187
Arith. Mean F-sc.	163	117	188	177
Harm. Mean F-sc.	163	117	188	174

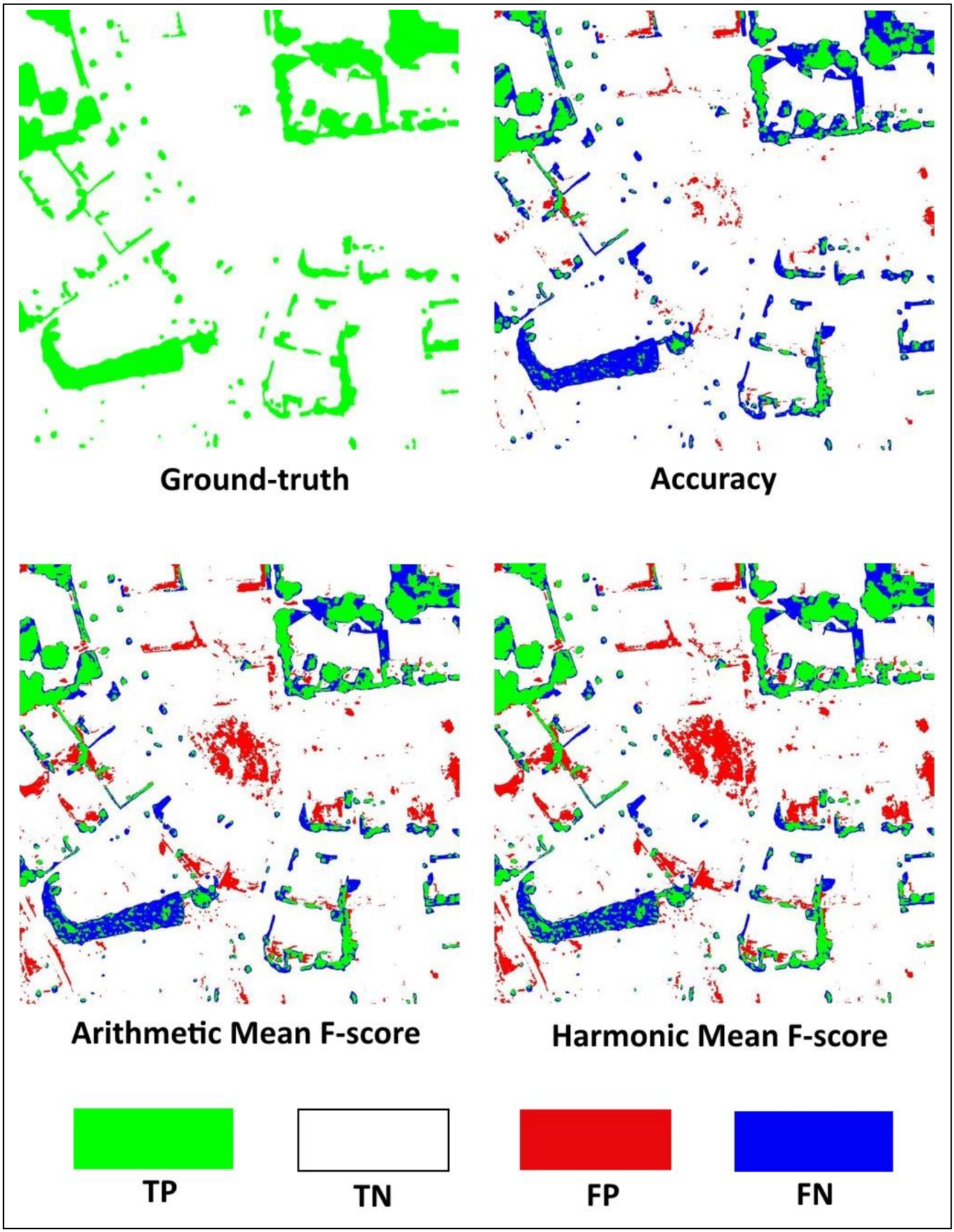


Figure 2: The comparison of the binary classification if the class "Tree" according to the different measurement methods.

In Figure 3, the distribution of the success of discrete threshold values according to three measurements methods is presented as graphically. Looking to the figures, it can be seen that the distribution of the harmonic mean F-score values provides a hill-climbing graphic required by the proposed method.

6 The continuous time threshold selection method and an experimental example

Instead of discretizing the values in the range 0-255 to integers and trying all of the integers as a threshold, the proposed method provides a continuous time convergence to the best threshold method by dividing the measurement distribution iteratively. The method needs a hill-climbing distribution, namely a decreasing distribution from the best threshold toward the two ends of the axis. Therefore, the distribution of the harmonic mean F-score measurement is the most suitable one among the three measurement methods for the continuous time threshold selection method proposed in this study. On the other hand, this method seems more reliable than the others; because this measurement gives the segmentation success as zero, when there is no segmentation where the threshold is one of the pool values (e.g. 0 and 255 for the range 0-255).

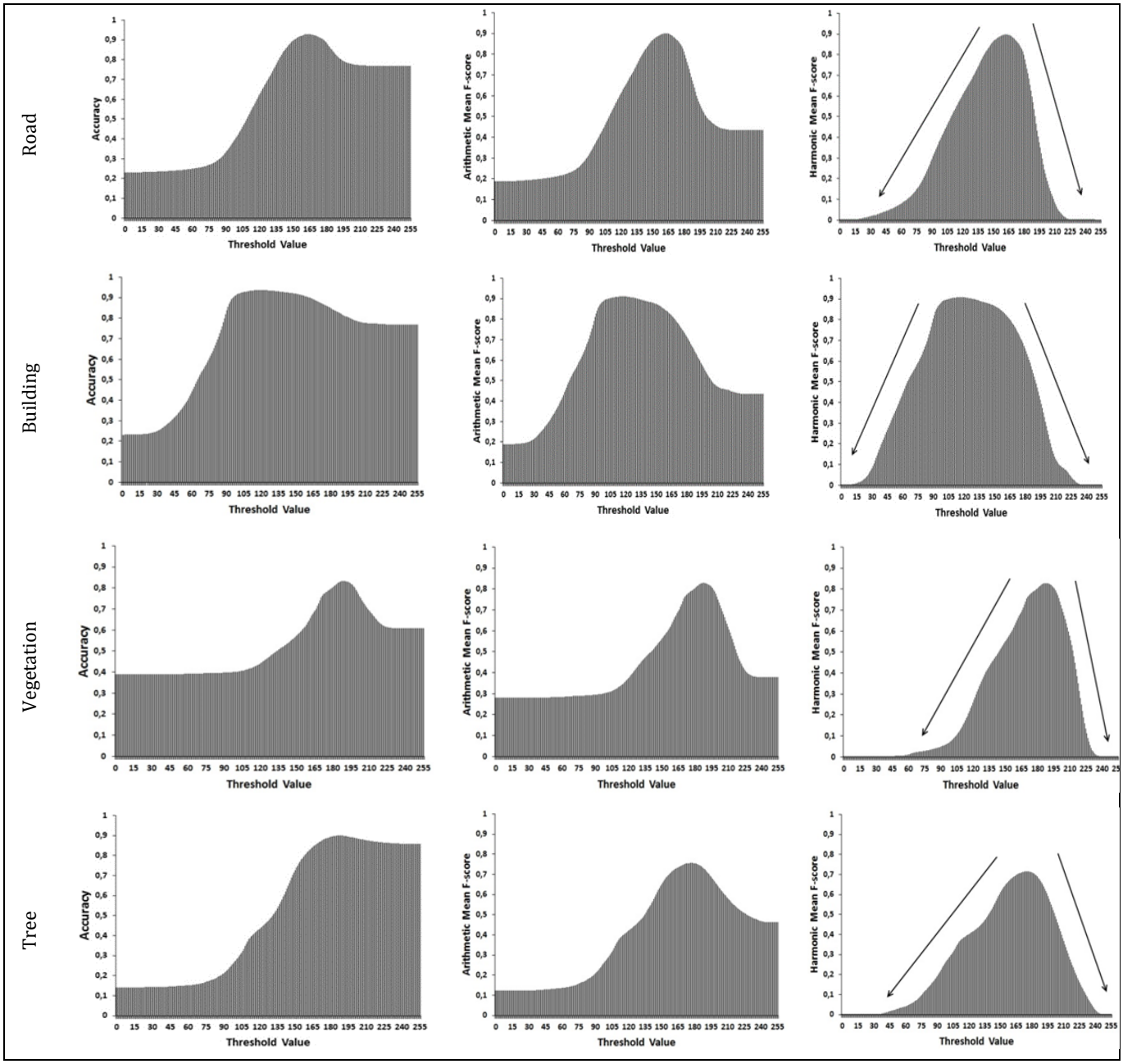


Figure 3. The distribution of the success of threshold values according to three measurements methods is presented as graphically (the arrows indicate that the distributions of harmonic mean F-scores are "hill-climbing").

According to the method, the range containing the best threshold value is divided to k fold equally. In this way, k+1 number of thresholds which represent the borderlines of the folds are added to an initially empty list. In first iteration, these thresholds are applied and their classification successes are measured. The least successful $k/2$ numbers of thresholds are eliminated from the threshold list and new $k/2$ numbers of thresholds in the middle of the consecutive threshold pairs of the remaining thresholds are added to the list. In the next iteration, the new threshold values are applied for the binary classification. When the iteration number reaches the predetermined threshold number or the score values of the consecutive threshold pairs converge each other as a specified tolerance value (a convergence degree), the algorithm ends. The threshold that gives the highest score in the last iteration is selected as the best threshold. After the first iteration, $k/2$ numbers of thresholds are applied to the data in each iteration. Determining the value of k is critical, because a very low value of k may ignore the best threshold over very peaky measurement distributions, while a very high value of k increases the execution time of the method.

In Figure 4 and Table 3, the method is applied for the class "Vegetation" onto the data using the harmonic mean F-score measurement where k=8. In Table 3, the thresholds which are added to the list for the first time in the iteration are colored with the same color.

7 Conclusion

In this study, we propose a continuous time threshold selection method for binary classification on polarized data to be used for released process on test data. The method converges to the best threshold as much as desired degree without being restricted to discrete values and runs in logarithmic time. We firstly applied the integers in the range 0-255 as discrete threshold values on a remote sensing image and examined their distributions of the success scores of the classification correctness measurement methods and the best threshold values for controlling the robustness of the proposed method (Figure 3 and Table 2). Looking the distributions of the measurement values, the values decrease from the best measurement value towards the worst measurement values. This property allows the proposed continuous time threshold selection method to converge to the best threshold. When we look as this aspect, the method needs data whose elements have discriminative features separating them from others, the ground-truth of the data and a data discrimination method such as LDA. The proposed method is iterative. The method adds new thresholds to be applied and eliminates the worst ones in each iteration such that the number of the added and removed thresholds is varying according to the number of initially applied thresholds. We applied the proposed method on the data for continuous time threshold value selection and checked the continuous time best threshold value on the discrete time best threshold values for verification. The method runs robustly in a logarithmic time with a suitable k parameter. In the feature studies, the method is thought to be part of a multi-class classification method.

8 References

[1] Lu D, Weng Q. "A survey of image classification methods and techniques for improving classification performance". *International Journal of Remote Sensing*, 28(5), 823-870, 2007.

[2] Wang W, Yang N, Zhang Y, Wang F, Cao T, Eklund P. "A review of road extraction from remote sensing images". *Journal of Traffic and Transportation Engineering*, 3(3), 271-282, 2016.

[3] Saglam A, Baykan NA. "A satellite image classification approach by using one dimensional discriminant analysis". *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3/W4, 429-435, 2018.

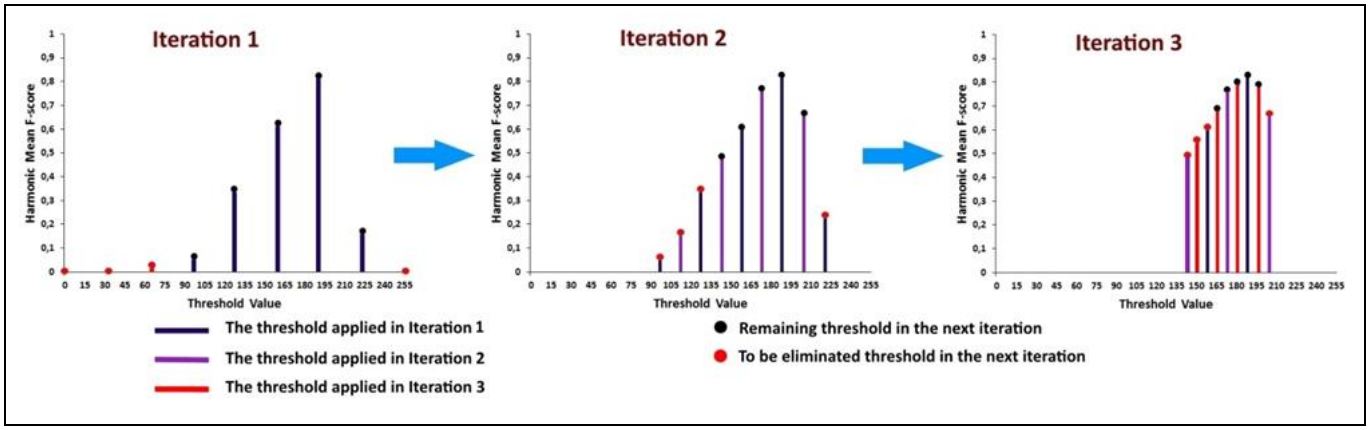


Figure 4. The first three iterations of the method for the class "Vegetation" using the harmonic mean F-score (k=8).

Table 3: The iterations (rows) and the tried thresholds with their classification score (columns) (threshold→score) for the class "Vegetation" using the harmonic mean F-score (k=8).

Iteration Number	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
1	1*→0	31.87*→0.002	63.75*→0.019	95.62→0.063	127.5→0.360	159.37→0.638	191.25**→0.822	223.12→0.171	255*→0
2	95.63*→0.063	111.56*→0.164	127.50*→0.360	143.44→0.499	159.38→0.638	175.31→0.786	191.25**→0.822	207.19→0.624	223.13*→0.171
3	143.44*→0.499	151.41*→0.564	159.38*→0.638	167.34→0.726	175.31→0.786	183.28→0.82	191.25**→0.822	199.22→0.761	207.19*→0.624
4	167.34*→0.726	171.330*→0.769	175.31*→0.786	179.30→0.802	183.28→0.82	187.27**→0.828	191.25→0.822	195.23→0.804	199.22*→0.761
5	179.30*→0.802	181.29*→0.811	183.28→0.82	185.27→0.826	187.27**→0.828	189.26→0.827	191.25→0.822	193.24*→0.815	195.23*→0.804

* : To be eliminated threshold in the next iteration,
**: The best threshold in the current iteration.

[4] Fisher RA. "The use of multiple measures in taxonomic problems". *Annals of Eugenics*, 7(2), 179-188, 1936.
[5] Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York, USA, Wiley, 2000.
[6] Martis RJ, Acharya UR, Min LC. "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform". *Biomedical Signal Processing and Control*, 8(5), 437-448, 2013.
[7] Lipton ZC, Elkan C, Naryanaswamy B. "Optimal thresholding of classifiers to maximize F1 measure". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8725(2), 225-239, 2014.
[8] Sanchez IE, Belgium A, Brun M. "Optimal threshold estimation for binary classifiers using game theory". *International Society for Computational Biology Community Journal*, 5(5), 1-11, 2016.
[9] Weinmann M, Jutzi B, Hinz S, Mallet C. "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers". *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286-304, 2015.
[10] Landrieu L, Raguet H, Vallet B, Mallet C, Weinmann M. "A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds". *ISPRS Journal of Photogrammetry and Remote Sensing*, 132, 102-118, 2017.
[11] Sokolova M, Lapalme G. "A systematic analysis of performance measures for classification tasks". *Information Processing and Management*, 45(4), 427-437, 2009.
[12] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. "Assessing the accuracy of prediction algorithms for classification: An overview". *Bioinformatics*, 16(5), 412-424, 2000.
[13] Freeman EA, Moisen GG. "A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa". *Ecological Modelling*, 217(1-2), 48-58, 2008.
[14] 2D Semantic Labeling Contest" "Online". http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (2018).
[15] Gerke M. "Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)". Department of Earth Observation Science, University of Twente, Enschede, Netherlands, Technical Report, 2015.
[16] Rottensteiner F, Sohn G, Gerke M, Baillard C, Benitez S, Breitkopf U. "ISPRS Test Project on Urban Classification and 3D Building Reconstruction". *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, I-3*, 293-298, 2012.