Journal of Universal Computer Science, vol. 26, no. 6 (2020), 720-733
submitted: 4/11/19, accepted: 3/4/20, appeared: 28/6/20

Integration of Decision Trees Using Distance to Centroid and to Decision Boundary

Jedrzej Biedrzycki and Robert Burduk
(Wroclaw University of Science and Technology
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
ORCID: 0000-0002-4924-1759, 0000-0002-3506-6611
{jedrzej.biedrzycki, robert.burduk}@pwr.edu.pl)

Abstract: Plethora of ensemble techniques have been implemented and studied in order to achieve better classification results than base classifiers. In this paper an algorithm for integration of decision trees is proposed, which means that homogeneous base classifiers will be used. The novelty of the presented approach is the usage of the simultaneous distance of the object from the decision boundary and the center of mass of objects belonging to one class label in order to determine the score functions of base classifiers. This means that the score function assigned to the class label by each classifier depends on the distance of the classified object from the decision boundary and from the centroid. The algorithm was evaluated using an open–source benchmarking dataset. The results indicate an improvement in the classification quality in comparison to the referential method – majority voting method.
Key Words: distance to decision boundary, classifier integration, ensemble of classifiers
Category: Topic I.5.2 - Design Methodology

1 Introduction

The supervised classification algorithm builds a mathematical model based on training data [Jordan and Mitchell, 2015]. This model is used to make predictions or decisions for a new object, in general, not belonging to the training set. Thus, the final effect of the recognition system uses the previously learned model to indicate the class label to the new object. In this general scenario, a classifier maps a feature space into a class label space. This mapping process can be decomposed into three stages. The first is to determine the value of the scoring function. The second one is the calibration of the scoring function, and the last is the conversion of the calibrated scoring function into a class label. For example, the scoring function of a linear SVM classifier is the object's distance from the decision boundary. Then, Platt scaling [Platt et al., 1999] computes the probability that a given object belongs to a particular class label.

The purpose of a classifier calibration is an approximation of the predicted scores to the actual probabilities. The calibration converts the scores function into probabilities, or more precisely transforms classifier outputs into values that can be interpreted as probabilities. The calibration methods can be generally divided into two groups: parametric and non-parametric methods. The sigmoidal transformation maps the score of a classifier to a calibrated probability output as was proposed by Platt [Platt et al., 1999]. The non-parametric methods are based on binning [Zadrozny and Elkan, 2001] or isotopic regression [Zadrozny and Elkan, 2002].

Ensemble methods (an ensemble of classifiers EoC) [Giacinto and Roli, 2001], [Ponti Jr, 2011], [Przybyła-Kasperek and Wakulicz-Deja, 2017] are a popular approach in building a classification model that is more stable and a model that uses a set of many individual classifiers (base learners) and combine them to classify new data [Le et al., 2013], [Rokach, 2010]. The main concept behind the ensemble technique is to create a classification method that outperforms every one of the base classifiers and, as it was previously mentioned, the outputs of the base classifiers can be used in a variety of ways to determine the decision of a classifiers committee.

The impact of distance from the decision boundary in boosting algorithms which are examples of EoC have been proposed. An emphasis function in which the first term takes large values for patterns with large quadratic error, and the second term increases for objects that lie close to the decision boundary is presented in paper [Gómez-Verdejo et al., 2010]. The emphasis function that balances also the contribution of the error and the distance to the decision boundary is considered in [Ahachad et al., 2017].

In this paper we present the concept of a scoring function which depends on the distance of the object from the decision boundary of a given base classifier and the centroid defined by the center of mass of objects belonging to one class label. The advantage of the proposed method is therefore the dependence of the scoring function on two coefficients: the distance from the decision boundary and the distance from the class label centroid. Experimental studies concern a classifiers committee built from heterogeneous base classifiers which are decision trees.

Given the above, the objectives of this work are the following:

– A proposal of a new score function that uses location of the cluster centroids defined by the class label and distance to the decision boundary defined by a base classifier.

– The use of the proposed score function in an ensemble of homogeneous decision tree classifiers.

– A new experimental setup to compare the proposed algorithm with the majority voting and random forest methods.

This paper is organized as follows: Section 2 presents the necessary terms of the classification. The proposed method for the calculation of the score function for decision trees is presented in Section 3. In the next sections experimental studies are discussed. Finally, some conclusions are presented.

2 Basic concept of classification

The recognition algorithm Ψ maps the feature space $\mathcal{X} \subseteq \Re^d$ to the set of class labels $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$ according to the general formula:

$$\Psi : \mathcal{X} \to \Omega. \tag{1}$$

For the feature vector $x \in \mathcal{X}$, that represents the recognised object Equation (1) can be written as

$$\Psi(x) = \omega_c. \tag{2}$$

The Equation (2) represents the so-called abstract level of the base classifier output [Kuncheva, 2014]. This level represents the information that the classifier Ψ assigns the unique class label ω_c to a given recognized object x, i.e. the output of the base classifier indicates uniquely the class label [Dey et al., 2014], [Przybyła-Kasperek and Wakulicz-Deja, 2017]. The other most commonly used type of the classifier output is the score function that addresses the degree of assigning the class label to the given recognized object x. An example of such a representation of the output is a posteriori probability returned by Bayes classifier [Bloch, 1996], [Ho et al., 1994].

Let us assume that K $(k \in \{1, 2, \ldots, K\})$ different decision trees $\Psi_1, \ldots, \Psi_K$ are used to solve the classification task. These decision trees are the base classifiers for the considered case of EoC. If all K base classifiers are equal contribution to make the final decision of EoC and the abstract level is considered, then the majority vote rule can be applied [Fechner and Keller, 2004], [Mohandes et al., 2018]. This method allows counting base classifiers outputs as a vote for a class and assigns the input pattern to the class with the greatest count of votes. It is defined as follows:

$$\Psi_{MV}(x) = \arg\max_{\omega_c} \sum_{k=1}^{L} I(\Psi_k(x) = \omega_c), \tag{3}$$

where $I(\cdot)$ is the indicator function. This function takes the value equal to 1 when the object described by the feature vector x is classified to the label ω_c, i.e. when $\Psi_k(x) = \omega_c$.

Another method of combining the base classifiers is the weighted voting. In this approach each of the base classifiers has an allocated weight, which may depend on the weight coefficient measured on the learning or validation dataset. In this approach Equation (3) takes the form:

$$\Psi_{MV}(x) = \arg\max_{\omega_c} \sum_{k=1}^{L} w_k * I(\Psi_k(x) = \omega_c), \tag{4}$$

where w_k is a weight of Ψ_k classifier.

If the output of each base classifier is a score function, Equation (3) can be expressed as

$$\Psi_{Sum}(x) = \arg\max_{\omega_c} \sum_{k=1}^{L} Sf_k(\omega_c), \tag{5}$$

where $Sf_k(\omega_c)$ is the score function of Ψ_k classifier calculated for ω_c class label. The weighting of Equation (6) takes a form:

$$\Psi_{Sum}(x) = \arg\max_{\omega_c} \sum_{k=1}^{L} w_k * Sf_k(\omega_c). \tag{6}$$

In the paper [Mao et al., 2015] this form of determining the output of EoC is called double weighting. This means that EoC takes into account the weight of each base classifier and the scoring function, which is treated as a certain weight.

In this article, we propose an algorithm in which weights are not assigned to base classifiers, but to an object that is recognised i.e. the weights are defined by the values of the scoring function.

3 Proposed method

Suppose that, the distance from the decision boundary is defined as the smallest distance from an object x_0 that would be assigned a different class by the base classifiers. The definition can be written formally using the following formula:

$$dist_B(x_0, \Psi) = \min_{x \in X; \Psi(x) \neq \Psi(x_0)} (dist(x; x_0)), \tag{7}$$

where X denotes the whole classification space, i.e. the cube based on feature values.

This is a general formula, that works for any classification algorithm. When the representation of the decision space cannot be described or is difficult to describe in terms of analytical functions, the decision space needs to be scanned in search for the solution. Of course, in the case of decision trees, for any given x_0, the closest point along feature axes needs to be found. The distance from the centroid is calculated for the centroid with the same label. The centroid's coordinates are calculated using the training subset.

Formally a score function for the given distance is defined as follows:

$$f(dist, \beta, \gamma) = \exp(-\gamma(dist - \beta)^2). \tag{8}$$

In this paper, β parameter of the scoring function for the distance from the decision boundary is fixed as $\beta_B = 0.5$ and for the distance from the centroid of the class label ω – as $\beta_\omega = 0$. This causes the scoring function to achieve its maximum at 0.5 and minimum at 0 and 1 when scoring the distance from

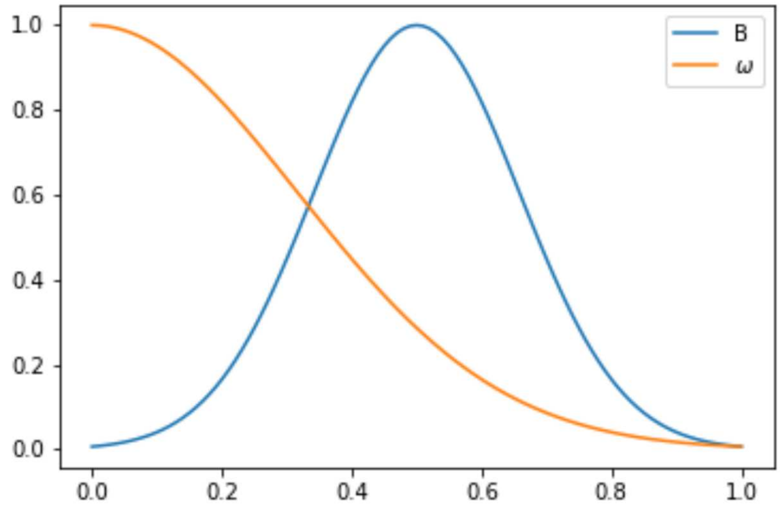

Figure 1: Score function for the boundary distance (B) and mass center distance (ω). Parameters: $\beta_B = 0.5$, $\beta_\omega = 0$, $\gamma_B = 20$, $\gamma_\omega = 5$.

decision boundary and maximum at 0 and minimum at 1 when scoring the distance from the centroid. Intuitively, this mean that the closer the object to the centroid, the larger the associated weight. Conversely, when the object is too close or too far from the decision boundary, the weight gets smaller. The combinations of γ parameter for the respective cases (γ_B and γ_ω) are depicted in Table 1. The base combination $\gamma_B = 20$ and $\gamma_\omega = 5$ is the base case, where the value of the function for the range of $[0, 1]$ varies between 0.0067 and 1 (so that the minimum is smaller than 1%). Additionally, cases where $\gamma_B = \gamma_\omega$ were examined. In our experiments the α parameter (the share of the distance from the decision boundary in weight computation) varies. The score functions are visualized in Fig. 3.

Finally, the scoring function is a linear combination of the scoring function for the decision boundary and centroid distances:

$$w_\Psi(x_0) = \alpha f\left(dist_B, \beta_B, \gamma_B\right) + (1-\alpha)\, f\left(dist_\omega, \beta_\omega, \gamma_\omega\right), \tag{9}$$

where $dist_B$ denotes the distance from the decision boundary defined by classifier Ψ and $dist_\omega$ – the distance from the centroid of class label ω. The parameter α is in the range $0 \leq \alpha \leq 1$.

Table 1: Combinations of γ parameter of the score function (8) examined.

γ_B	γ_ω
20	5
5	5
20	20
10	10

Formally, the proposed classification algorithm can be presented as:

$$\Psi_\alpha(x, \beta_B, \beta_\omega, \gamma_B, \gamma_\omega) = \arg\max_{\omega_c} \sum_{k=1}^{K} I(\Psi_k(x), \omega_c) w_{\Psi_k}(x) \qquad (10)$$

The pseudocode of the proposed approach to creating EoC with score function depends on two coefficients in geometric space is given in Algorithm 1.

Algorithm 1: Classification algorithm based on distance from decision boundary and centroid of the class label.

Input : K – number of base classifiers $(\Psi_1, \Psi_2, \ldots, \Psi_K)$, α - contribution of the distance from the decision boundary to weight calculation $(0 \leq \alpha \leq 1)$, x – the classified object

Output: ω_c – the label predicted by the integrated classifier

Scale all features into the range $[0, 1]$.

Split the dataset into $K + 1$ subsets (K for training every base decision tree and 1 for testing purposes).

for $k := 1$ ***to*** K **do**

  Determine centroids for each class label using learning subsets.

  Train a base classifier Ψ_k using k-th learning subset

  Calculate $w_{\Psi_k}(x)$

end

return *Output the final decision of the ensemble classifier:* $\Psi_x = \arg\max_{\omega_c} \sum_{k=1}^{K} I(\Psi_k(x), \omega_c) w_{\Psi_k}(x)$

---

4 Experimental Setup

In the experiment decision trees as base classifiers were used and a pool of classifiers consisting of five decision trees was created, i.e. $K = 5$ and EoC consists of five base classifiers. The decision tree implementation from apache spark was utilised. The integration algorithm was implemented in scala. To perform statistical tests Numpy [Oliphant, 2020] and scipy [Jones et al., 2020] were used.

The experiments were conducted using open-source data sets available on platforms UCI Machine Learning Repository [Dua and Graff, 2017] and KEEL Data Set Repository [Alcalá-Fdez et al., 2011]. For clarity the following abbreviations for datasets names are used: bio – Biodeg, bup – Bupa, cry – Cryotherapy, dba – Data banknote authentication, hab – Haberman, ion – Ionosphere, met – Ultrasonic flowmeter diagnostics, pop – Pop failures, sei – Seismic bumps, wdb – Breast Cancer Wisconsin (Diagnostic), wis – Breast Cancer Wisconsin (Original).

For all datasets the feature selection process was performed to indicate two most informative features [Guyon and Elisseeff, 2003], [Rejer, 2015]. In the case of the two-dimensional space classification, decision trees can be considered in the geometric space as finite sets of rectangular regions with a specified class label.

Table 2: ACC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 20$ and $\gamma_\omega = 5$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.715	0.658	0.742	0.915	0.717	0.736	0.777	0.915	0.938	0.906	0.955	2.73
$\Psi_{0.3}$	0.722	0.497	0.761	0.907	0.691	0.789	0.762	0.914	0.935	0.889	0.937	3.68
$\Psi_{0.7}$	0.726	0.528	0.769	0.907	0.725	0.730	0.763	0.902	0.926	0.907	0.953	3.68
$\Psi_{1.0}$	0.692	0.582	0.758	0.921	0.742	0.746	0.713	0.913	0.927	0.888	0.929	4.09
Ψ_{mv}	0.720	0.568	0.716	0.912	0.707	0.759	0.753	0.910	0.931	0.900	0.935	4.23
Ψ_{rf}	0.724	0.546	0.840	0.919	0.746	0.775	0.727	0.911	0.931	0.909	0.944	2.59

Table 3: MCC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 20$ and $\gamma_\omega = 5$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.435	0.304	0.514	0.828	0.171	0.417	0.484	0.000	0.000	0.801	0.896	2.73
$\Psi_{0.3}$	0.377	-0.026	0.605	0.815	0.072	0.460	0.501	0.000	-0.003	0.773	0.859	4.09
$\Psi_{0.7}$	0.405	0.059	0.589	0.817	0.023	0.491	0.524	0.000	0.000	0.804	0.890	3.18
$\Psi_{1.0}$	0.236	-0.067	0.505	0.843	0.072	0.445	0.437	0.000	0.000	0.767	0.841	4.59
Ψ_{mv}	0.406	0.095	0.426	0.826	0.032	0.465	0.488	-0.002	0.022	0.791	0.853	3.82
Ψ_{rf}	0.415	0.078	0.691	0.836	0.169	0.512	0.468	0.000	-0.003	0.810	0.876	2.59

Table 4: ACC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 5$ and $\gamma_\omega = 5$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.729	0.624	0.806	0.924	0.637	0.792	0.701	0.917	0.930	0.918	0.961	2.55
$\Psi_{0.3}$	0.685	0.593	0.800	0.910	0.723	0.769	0.624	0.899	0.933	0.901	0.942	4.41
$\Psi_{0.7}$	0.718	0.608	0.680	0.915	0.756	0.779	0.750	0.931	0.933	0.906	0.954	2.86
$\Psi_{1.0}$	0.737	0.494	0.738	0.917	0.768	0.710	0.683	0.926	0.935	0.864	0.928	3.73
Ψ_{mv}	0.720	0.568	0.716	0.912	0.707	0.759	0.753	0.910	0.931	0.900	0.935	4.41
Ψ_{rf}	0.724	0.546	0.840	0.919	0.746	0.775	0.727	0.911	0.931	0.909	0.944	3.05

Table 5: MCC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 5$ and $\gamma_\omega = 5$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.412	0.175	0.614	0.846	0.027	0.500	0.387	0.000	-0.007	0.827	0.917	2.64
$\Psi_{0.3}$	0.321	0.145	0.587	0.823	0.093	0.496	0.216	0.000	-0.005	0.790	0.873	4.27
$\Psi_{0.7}$	0.403	0.113	0.323	0.830	0.193	0.490	0.348	0.000	-0.003	0.793	0.899	3.59
$\Psi_{1.0}$	0.449	-0.010	0.477	0.838	0.126	0.428	0.378	0.000	0.000	0.710	0.831	3.91
Ψ_{mv}	0.406	0.095	0.426	0.826	0.032	0.465	0.488	-0.002	0.022	0.791	0.853	4.09
Ψ_{rf}	0.415	0.078	0.691	0.836	0.169	0.512	0.468	0.000	-0.003	0.810	0.876	2.50

Table 6: ACC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 20$ and $\gamma_\omega = 20$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.705	0.623	0.746	0.907	0.712	0.779	0.732	0.892	0.941	0.907	0.956	3.00
$\Psi_{0.3}$	0.711	0.571	0.881	0.907	0.694	0.794	0.671	0.878	0.929	0.899	0.959	3.77
$\Psi_{0.7}$	0.698	0.520	0.721	0.903	0.712	0.771	0.633	0.914	0.934	0.893	0.928	4.86
$\Psi_{1.0}$	0.702	0.573	0.740	0.911	0.735	0.787	0.747	0.903	0.937	0.914	0.931	2.91
Ψ_{mv}	0.720	0.568	0.716	0.912	0.707	0.759	0.753	0.910	0.931	0.900	0.935	3.77
Ψ_{rf}	0.724	0.546	0.840	0.919	0.746	0.775	0.727	0.911	0.931	0.909	0.944	2.68

To evaluate the proposed methods the following classification measures are used: average accuracy (ACC) and Matthews correlation coefficient (MCC).

Table 7: MCC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 20$ and $\gamma_\omega = 20$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.402	0.173	0.482	0.816	0.205	0.490	0.454	-0.027	0.033	0.801	0.897	3.00
$\Psi_{0.3}$	0.408	0.081	0.768	0.811	0.111	0.530	0.367	0.000	-0.003	0.790	0.910	3.09
$\Psi_{0.7}$	0.327	-0.039	0.448	0.804	0.083	0.460	0.282	0.000	0.000	0.777	0.841	5.09
$\Psi_{1.0}$	0.203	0.060	0.470	0.822	0.030	0.518	0.467	0.000	0.000	0.817	0.848	3.73
Ψ_{mv}	0.406	0.095	0.426	0.826	0.032	0.465	0.488	-0.002	0.022	0.791	0.853	3.55
Ψ_{rf}	0.415	0.078	0.691	0.836	0.169	0.512	0.468	0.000	-0.003	0.810	0.876	2.55

Table 8: ACC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 10$ and $\gamma_\omega = 10$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.746	0.632	0.745	0.915	0.730	0.752	0.779	0.907	0.931	0.904	0.925	3.14
$\Psi_{0.3}$	0.742	0.608	0.768	0.903	0.707	0.758	0.606	0.917	0.942	0.919	0.932	3.14
$\Psi_{0.7}$	0.734	0.528	0.848	0.914	0.693	0.781	0.643	0.889	0.928	0.895	0.925	4.32
$\Psi_{1.0}$	0.710	0.530	0.594	0.911	0.715	0.769	0.718	0.914	0.936	0.915	0.924	4.00
Ψ_{mv}	0.720	0.568	0.716	0.912	0.707	0.759	0.753	0.910	0.931	0.900	0.935	3.86
Ψ_{rf}	0.724	0.546	0.840	0.919	0.746	0.775	0.727	0.911	0.931	0.909	0.944	2.55

Table 9: MCC values and Friedman rank of integrated classifiers and random forest for $\gamma_B = 10$ and $\gamma_\omega = 10$.

	bio	bup	cry	dba	hab	ion	met	pop	sei	wdb	wis	rank
$\Psi_{0.0}$	0.434	0.310	0.508	0.829	0.231	0.497	0.592	-0.024	-0.004	0.797	0.825	3.59
$\Psi_{0.3}$	0.464	0.193	0.574	0.804	0.032	0.503	0.209	0.000	-0.003	0.827	0.845	3.41
$\Psi_{0.7}$	0.435	0.054	0.705	0.829	-0.017	0.544	0.294	0.000	0.000	0.785	0.839	4.59
$\Psi_{1.0}$	0.263	0.059	0.358	0.821	0.074	0.515	0.416	0.000	0.000	0.816	0.841	3.82
Ψ_{mv}	0.406	0.095	0.426	0.826	0.032	0.465	0.488	-0.002	0.022	0.791	0.853	3.86
Ψ_{rf}	0.415	0.078	0.691	0.836	0.169	0.512	0.468	0.000	-0.003	0.810	0.876	2.73

MCC is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories [Chicco and Jurman, 2020].

As reference classifiers the majority voting Ψ_{mv} and random forest Ψ_{rf} EoC were used.

5 Results and Discussion

The main aim of the experiments was to compare the quality of the classification of the proposed method of base classifiers integration in the geometric space with random forest (Ψ_{rf}) and majority voting (Ψ_{mv}). The integrated classifier with the given α is denoted as Ψ_{α}. In order to compare the quality of the classification, two classification measures: accuracy (ACC) and Matthews correlation coefficient (MCC) were used. Tab. 2, 4, 6, 8 shows the results of ACC and Tab. 3, 5, 7, 9 shows the results of MCC. Along with the quality measures, average ranks obtained in nonparametric Friedman tests [Demšar, 2006], [Trawinski et al., 2012] are presented in the last column. Four different values of α (share of distance from the decision boundary) were examined: $\alpha \in \{0, 0.3, 0.7, 1\}$.

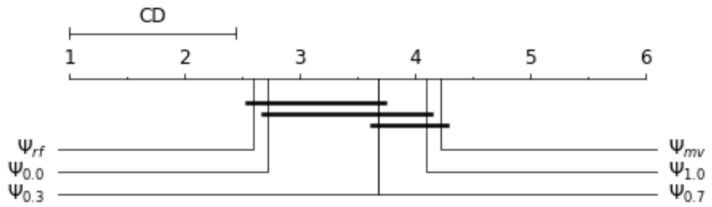

Figure 2: Friedman ranks comparison for ACC measure $\gamma_B = 20$ and $\gamma_\omega = 5$.

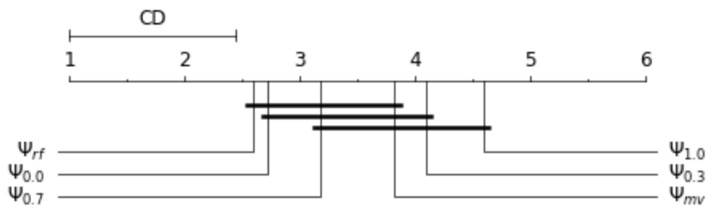

Figure 3: Friedman ranks comparison for MCC measure $\gamma_B = 20$ and $\gamma_\omega = 5$.

Bonferroni–Dunn tests were conducted to find the difference between the examined algorithms. For the significance level of 0.1 the critical value of difference between Friedman ranks for the test equals 1.44 (5 algorithms are compared against the reference method, 11 data sets are used). The results of the statistical

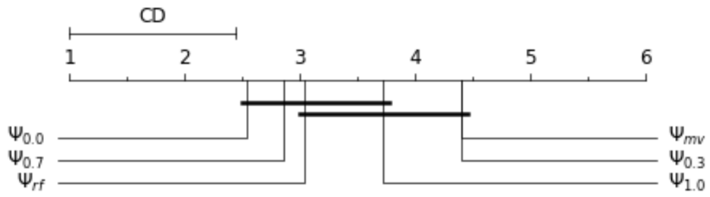

Figure 4: Friedman ranks comparison for ACC measure $\gamma_B = 5$ and $\gamma_\omega = 5$.

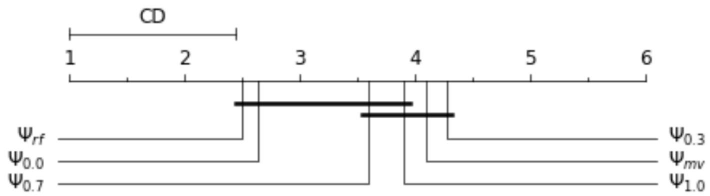

Figure 5: Friedman ranks comparison for MCC measure $\gamma_B = 5$ and $\gamma_\omega = 5$.

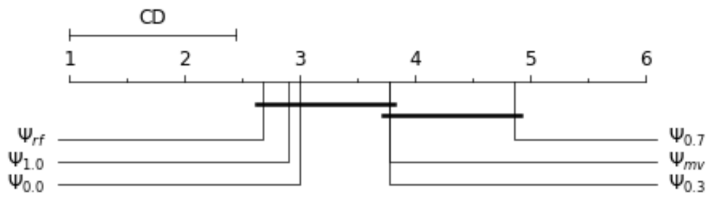

Figure 6: Friedman ranks comparison for ACC measure $\gamma_B = 20$ and $\gamma_\omega = 20$.

analysis are presented in Fig. 2–9. In the conducted experiments the integration algorithm outperformed the random forest only in case of ACC for $\alpha \in 0, 0.7$ and $\gamma_B = \gamma_\omega = 5$ although changes are not significant. Integrated classifier created with parameter $\alpha = 0$ yields significantly better results than majority voting for $\gamma_B = \gamma_\omega = 5$ (both ACC and MCC) and $\gamma_B = 20, \gamma_\omega = 5$ (ACC). Additionally for $\gamma_B = \gamma_\omega = 5$, ACC of integrated classifier is significantly better for $\alpha = 0.7$ than ACC of majority voting.

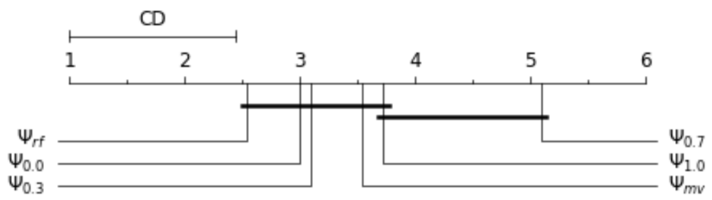

Figure 7: Friedman ranks comparison for MCC measure $\gamma_B = 20$ and $\gamma_\omega = 20$.

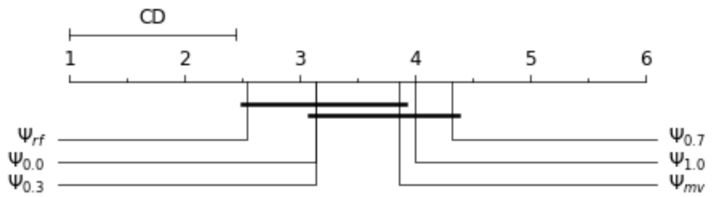

Figure 8: Friedman ranks comparison for ACC measure $\gamma_B = 10$ and $\gamma_\omega = 10$.

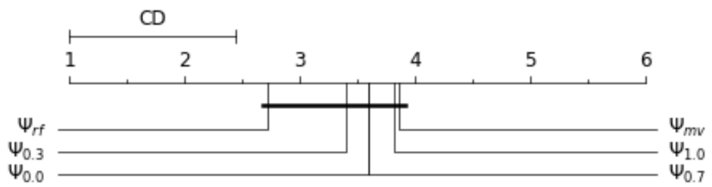

Figure 9: Friedman ranks comparison for MCC measure $\gamma_B = 10$ and $\gamma_\omega = 10$.

6 Conclusion

In this article the algorithm of decision tree integration in the geometric space was proposed. In detail we propose the score function that takes into account the distance of recognized object from the centroid and the distance from the decision boundary. Eleven open-source benchmarking datasets were used in the experimental part to perform the statistical analysis of results concerning two classification measures ACC and MCC. Bonferroni–Dunn test showed, that, in most cases, the proposed algorithm for $\alpha = 0$ performs better than majority voting. This means that the proposed method is more effective than majority voting. Additionally obtained results indicate that the distance from the centroid of class label is more significant than the distance from the decision boundary in case of EoC consisting of decision trees.

Acknowledgments

This work was supported by the National Science Centre, Poland under the grant no. 2017/25/B/ST6/01750.

References

[Ahachad et al., 2017] Ahachad, A., Álvarez Pérez, L., and Figueiras-Vidal, A. R. (2017). Boosting ensembles with controlled emphasis intensity. *Pattern Recognit. Lett.*, 88:1–5.

[Alcalá-Fdez et al., 2011] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

[Bloch, 1996] Bloch, I. (1996). Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on systems, man, and cybernetics-Part A: systems and humans*, 26(1):52–67.

[Chicco and Jurman, 2020] Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

[Dey et al., 2014] Dey, A., Shaikh, S. H., Saeed, K., and Chaki, N. (2014). Modified majority voting algorithm towards creating reference image for binarization. In *Advanced Computing, Networking and Informatics-Volume 1*, pages 221–227. Springer.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Fechner and Keller, 2004] Fechner, B. and Keller, K. (2004). A fault-tolerant voting scheme for multithreaded environments. In *Parallel Computing in Electrical Engineering, 2004. International Conference on*, pages 237–239. IEEE.

[Giacinto and Roli, 2001] Giacinto, G. and Roli, F. (2001). An approach to the automatic design of multiple classifier systems. *Pattern recognition letters*, 22(1):25–33.

[Gómez-Verdejo et al., 2010] Gómez-Verdejo, V., Arenas-García, J., and Figueiras-Vidal, A. R. (2010). Committees of adaboost ensembles with modified emphasis functions. *Neurocomputing*, 73(7-9):1289–1292.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. e. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

[Ho et al., 1994] Ho, T. K., Hull, J. J., and Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence*, 16(1):66–75.

[Jones et al., 2020] Jones, E., Oliphant, T., Peterson, P., et al. (2020). SciPy : Open source scientific tools for Python . [Online: accessed 10.02.2020].

[Jordan and Mitchell, 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[Kuncheva, 2014] Kuncheva, L. I. (2014). *Combining Pattern Classifiers*. John Wiley & Sons, Inc.

[Le et al., 2013] Le, T. N., Ho, T. B., Kawasaki, S., Kanda, T., Takabayashi, K., Wu, S., and Yokosuka, O. (2013). A semi-supervised ensemble learning method for finding discriminative motifs and its application. *Journal of Universal Computer Science*, 19(4):563–580.

[Mao et al., 2015] Mao, S., Jiao, L., Xiong, L., Gou, S., Chen, B., and Yeung, S.-K. (2015). Weighted classifier ensemble based on quadratic form. *Pattern Recognition*, 48(5):1688–1706.

[Mohandes et al., 2018] Mohandes, M., Deriche, M., and Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639.

[Oliphant, 2020] Oliphant, T. (2020). NumPy : A guide to NumPy . USA: Trelgol Publishing. [Online: accessed 10.02.2020].

[Platt et al., 1999] Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[Ponti Jr, 2011] Ponti Jr, M. P. (2011). Combining classifiers: from the creation of ensembles to the decision fusion. In *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, pages 1–10. IEEE.

[Przybyła-Kasperek and Wakulicz-Deja, 2017] Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2017). Dispersed decision-making system with fusion methods from the rank level and the measurement level – a comparative study. *Information Systems*, 69:124–154.

[Przybyła-Kasperek and Wakulicz-Deja, 2017] Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2017). Comparison of fusion methods from the abstract level and the rank level in a dispersed decision-making system. *International Journal of General Systems*, 46(4):386–413.

[Rejer, 2015] Rejer, I. (2015). Genetic algorithms for feature selection for brain-computer interface. *IJPRAI*, 29.

[Rokach, 2010] Rokach, L. (2010). *Pattern classification using ensemble methods*, volume 75. World Scientific.

[Trawinski et al., 2012] Trawinski, B., Smetek, M., Telec, Z., and Lasota, T. (2012). Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22(4):867–881.

[Zadrozny and Elkan, 2001] Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*. ACM Press.