Abstract

Popularity of artificial intelligence and machine learning methods is still growing and can be found in fields like cybersecurity, optimization, finance, medicine and healthcare, law, media or education.

Supervised learning is one of the methods of machine learning. The classification algorithms build a model, also called a classifier, with the use of labelled data building up the training set. The created model is then utilized to label the new, not yet labelled objects.

An important role among classification algorithms play ensemble classifiers. It was noticed that utilizing multiple models forming a system leads to improvement in classification performance.

Research problem

One such approach to classifier integration is by using their geometric representation expressed by the decision boundaries. Utilizing decision boundaries, which are the result of classifier training, allows for defining novel algorithms for model integration. These algorithms do not use the labels' probability vector or the labels themself.

In the dissertation, four novel algorithms using decision boundaries produced by decision trees are proposed.

The main goal of the study was to broaden the knowledge of diverse multiclassifier system creation methods based on their geometric features, proposal of new algorithms and their implementation as well as evaluation against referential techniques: majority voting and random forest. The research hypothesis of this dissertation is formulated as follows.

Research hypothesis. *Utilization of trained decision trees' decision boundaries allows for building an ensemble of classifiers with a greater value of performance quality measure than the multiclassifier system like random forest or majority voting using the same set of trained decision trees.*

Designed algorithms of classifier ensembling were compared to other ensemble techniques: (weighted) majority voting and random forest. The following classification performance measures were used to prove the research hypothesis: accuracy (ACC), Matthews Correlation Coefficient (MCC) and F-score for binary classification datasets and accuracy, precision, recall and F-score (three latter both micro– and macroaveraged) for the other problems. The experimental studies were performed using testing, open–source datasets from UCI and KEEL platforms. The proposed research hypothesis was verified by non–parametric statistical tests involving diverse classification quality measures.

Four algorithms for decision tree integration employing their geometric representation were developed, implemented and evaluated. Statistical analysis of the results obtained for the datasets mentioned indicates that the proposed methods outperform in many cases the referential integration techniques. In the dissertation it was proven, that the proposed algorithm is a generalization of weighted majority voting.

Achieved results

The first integration algorithm utilizes two distances in the feature space used as the coordinate system with the classification boundaries of trained classifiers projected. The first distance is calculated from the centroid and the second – from the decision boundary. The object under test is classified by the model of the trained decision tree itself. The centroid of the corresponding class is taken to compute the distance. Both distances were mapped with the Gaussian function. Several parameter values of the mapping function were studied. The additional mapping serves two purposes:

- The distances are normalized to the unit range: $[0, 1]$.
- The calculated value reflects the contribution of the distance from the decision boundary to the weight value.

The final decision was computed as the class label for which the linear combination of both distances is maximal. Several possible distribution parameters were examined in order to find the highest quality measure. The statistical tests indicated the significant

improvement in the classification quality measures: ACC and MCC of the presented technique in comparison with the referential ensemble methods.

The second proposal is an algorithm where the ensemble model is produced using the static space division. The feature space is split into equal subspaces, with the same number of divisions along every axis, whose label is determined by the classification of the middle point. For the particular label, the weight is computed for the subspaces based on the volume of classification regions. Classification regions are the cuboids with the highest volume spanning over the points of a single decision tree classification. Additional conference article presents a theoretical proof that the presented algorithm is the generalization of the weighted majority voting. The implementation was examined using the datasets. Statistically significant classification quality improvement was found, especially in the quality measures like ACC and MCC.

In the third method proposed, dynamic space partitioning was employed. The geometric representation of the trained model determines the division of the feature space. For every label, the weight is calculated based on the classification of the region itself and its neighbors. Half of the weight is assigned from the subspace itself and the weights of its neighbors sum up to the other half. The weights depend on the distances between the middle points of the neighboring subspaces. The statistical analysis was conducted and the performance improvement expressed with microaveraged precision, recall and F-score in comparison with the referential majority voting and random forest was found.

In the last work, similarly as in the previous one, dynamic division of feature space is applied and the subspaces' classification is influenced by the surrounding subspaces. However, additional improvements were introduced:

- Mean of the training points is used instead of the geometric middle point of the subspace in the definition of the distance between the regions. The distance between the averages is taken as the distance between the subspaces.
- Only the subspaces containing the objects among which the ones with the minor label made up at least 5% of all the objects in that subspace were taken during weight calculation.
- A wider range of neighbor subspaces is taken in case the filtering rejects the whole neighbor ring. The procedure is conducted until a non–empty ring is found.

The discussed algorithm was tested using the datasets mentioned against both referential methods (majority voting and random forest) and the previously presented, original algorithm. Statistically significant improvement is observed in the quality performance against all the reference methods. In particular for the following measures: accuracy,

micro– and macro–averaged F-score statistically significant difference was found, which proves the research hypothesis.

A consistent experimental setup is followed in all the studies. Comprehensive evaluation on multiple benchmarking datasets was conducted. Several classification quality measures are computed to account for diverse datasets' features, like imbalance ratio: ACC, MCC and F-score for binary and micro– and macro–averaged precision, recall and F-score for other problems. Spark with Scala was harnessed to implement every presented algorithm. The source code is hosted and publicly available on github for other researchers working on machine learning problems to verify: https://github.com/TAndronicus/dtree-merge, https://github.com/TAndronicus/dtree-merge-scoring, https://github.com/TAndronicus/dynamic-dtree, https://github.com/TAndronicus/dynamic-ring.

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