Early and Late Fusion of Classifiers for the MediaEval Medico Task

Mario Taschwer[1], Manfred Jürgen Primus[1], Klaus Schoeffmann[1], Oge Marques[2]
[1]Klagenfurt University (AAU), Austria, [2]Florida Atlantic University (FAU), USA

ABSTRACT

In this paper we present our results for the MediaEval 2018 Medico task, achieved with traditional machine learning methods, such as logistic regression, support vector machines, and random forests. Before classification, we combine traditional global image features and CNN-based features (early fusion), and apply soft voting for combining the output of multiple classifiers (late fusion). Linear support vector machines turn out to provide both good classification performance and low run-time complexity for this task.

1 INTRODUCTION

The *Medico* task at MediaEval 2018 [9] addresses the problem of predicting a predefined set of diseases and findings in endoscopic images of the gastrointestinal (GI) tract of the human body. Participating teams are provided with a development set of 5293 images labeled with 16 different classes by medical experts specialized in GI inspection. The development set has been sampled from Kvasir [8] and Nerthus [7] video datasets. Teams are supposed to develop classifiers that are able to predict these classes on unseen images with low run time complexity. Task organizers evaluate and compare submitted approaches based on two main measures: (1) correlation between predictions on the test set and ground truth, using Matthews correlation coefficient (MCC) [1], and (2) required processing time of predictions. The imbalance and sparsity of the training set (see Fig. 1) poses a particular challenge of this task.

2 APPROACH

To address the research objectives described in Section 1, we chose to train traditional machine learning algorithms only [3, 12], but combine them with each other and with CNN-based feature extraction using a two-level fusion strategy: (1) *early fusion* of feature vectors by concatenating them; and (2) *late fusion* of classifiers by averaging their predicted class probabilities.

Early fusion of feature vectors created by different feature extraction methods increases the dimensionality of the feature space and improves the likelihood that binary classification problems are linearly separable. On the other hand, increasing feature dimensionality will reduce the run-time efficiency of machine learning algorithms. To cope with this trade-off, we selected a sensible combination taken from the following feature sets of the given endoscopic image dataset: *LIRE* – traditional global image features extracted using the LIRE library [4], as provided by Medico task organizers (1185-dimensional after concatenation); *GoogLeNet* – output of the last hidden layer of GoogLeNet CNN [11], trained on ImageNet (1024-dimensional); *SurgicalAction* – output of the last hidden layer of GoogLeNet CNN, trained on a dataset of laparoscopic surgery videos for detecting surgical actions in gynecologic laparoscopy [5] (1024-dimensional).

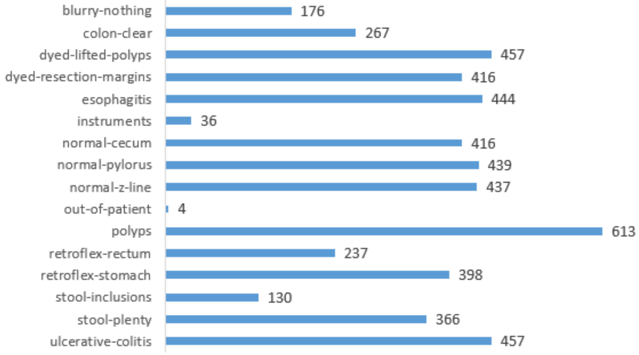

Figure 1: Number of training examples per class

Preliminary cross-validation experiments on the training dataset indicated that *LIRE* and *GoogLeNet* features performed significantly better than *SurgicalAction* features, consistently across different classifiers. Moreover, early fusion (concatenation) of *LIRE* and *GoogLeNet* features (2209-dimensional) further improved cross-validation scores, but no additional improvement was observed when combining all three feature sets. Therefore, we decided to use only the 2209-dimensional feature set for all further experiments producing detection runs for the Medico task. Each feature dimension was centered (by subtracting the mean) and scaled to the range $[-1, 1]$ before fed into classifiers.

Due to the low number of training samples (see Fig. 1) compared to the dimensionality of the feature space, traditional linear classifiers have the potential to provide a sensible trade-off between effectiveness (classification performance) and efficiency (run-time complexity). We therefore included two linear classifiers, logistic regression (LR) and linear support vector machine (LSVM), into our experiments, but also chose two non-linear classifiers for comparison, namely random forests (RF) and kernel support-vector machine (KSVM) with a radial basis function (RBF) kernel. Multinomial logistic regression and random forests implicitly support multi-class classification, whereas support-vector machines (SVMs) were used in a one-versus-rest (OVR) ensemble to support multiple classes.

Table 1 lists the classifiers used for our submitted runs, including information on decision boundaries, tuned hyper-parameters, multi-class strategy, support for class probabilities, and obtained cross-validation (CV) scores on the development set. Hyper-parameters include the regularization parameter C (lower values mean stronger regularization), the width γ of the RBF kernel, and number n and maximal depth k of decision trees for random forests.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Table 1: Classifiers used for submitted runs

Classifier	Boundary	Parameters	Multi-class	Probabilities	MCC_CV
LR	linear	$C = 0.1$	implicit	yes	0.8677
LSVM	linear	$C = 0.01$	OVR	no	**0.8699**
KSVM	non-linear	$C = 10, \gamma = 0.001$	OVR	yes*	0.8673
RF	non-linear	$n = 500, k = 30$	implicit	yes	0.8444

* class probabilities computed by Platt scaling

Hyper-parameters of chosen classifiers were optimized independently using grid search in the parameter space and 4-fold cross-validation on the development set provided by task organizers. As objective function for optimization the mean MCC [1] was used. After hyper-parameter selection, final classification models were trained on the entire development dataset, without using additional training data.

To combine the output of several classifiers (late fusion), various well-known ensemble methods exist in the literature [10]. If classification performances of component classifiers are similar (as for the classifiers listed in Table 1), a simple approach to late fusion, called *soft voting*, is often effective. For a given test instance, soft voting computes average probabilities for each class over all component classifiers and finally predicts the class with maximal average probability.

Note that only two of the chosen classifiers naturally provide class probabilities for prediction (LR and RF), whereas SVMs do not. However, with additional run-time cost, confidence scores produced by SVMs can be transformed into class probabilities using Platt scaling [6]. Since we used the Scikit-learn Python framework [3] to perform classification experiments and Platt scaling was implemented only for KSVM but not for LSVM, we considered only three of the chosen classifiers (LR, RF, KSVM) for late fusion experiments.

3 RESULTS

We submitted five runs with predictions for the test set (8740 images) to Medico task organizers for evaluation: one for each classifier listed in Table 1, and a late fusion run combining the output of RF, KSVM, and LR classifiers. Table 2 lists some of the evaluation metrics obtained by official evaluation of our runs as well as mean prediction times per image (in milliseconds) on the test set measured on commodity PC hardware (Intel Core2 E8400 CPU @ 3 GHz, 8 GB RAM, no GPU usage). Note that prediction times do not include feature extraction and model loading times, but refer to the time span needed to perform feature scaling and prediction of classes and class probabilities (if applicable), including the total time needed to apply an ensemble of classifiers (RF, SVM in OVR mode, RF-KSVM-LR fusion). Prediction times have been measured three times and the average value is reported in the table. Since feature extraction and prediction were performed with different software and in batch mode, investigation of real-time processing capabilities of the proposed approach for online video processing would need further experiments.

Table 2: Evaluation results of submitted runs, including mean prediction time T per image

Run	accuracy	F1	MCC	T / ms
LR	0.9873	0.8986	0.8919	0.119
LSVM	**0.9876**	**0.9008**	**0.8942**	**0.103**
KSVM	0.9865	0.8921	0.8849	25.808
RF	0.9843	0.8747	0.8664	0.828
RF-KSVM-LR	0.9875	0.9002	0.8936	26.783

4 DISCUSSION AND CONCLUSIONS

All submitted runs display comparable classification performances on the test set, except for RF, which shows a slightly degraded performance. We explain this by a possible over-regularization due to limiting the maximal tree depth to 30 during training.

Remarkably, LSVM displays a slight advantage with respect to classification performance in comparison to all other runs, including the late fusion approach (RF-KSVM-LR). Moreover, the slight advantage of LSVM turns into a substantial gain when taking also run-time efficiency into account. The high computational costs of KSVM and of RF-KSVM-LR are mostly due to expensive Platt scaling.

When comparing test results to cross-validation scores on the development set (see Table 1), it may come as a surprise that classification performance has improved on the test set. We explain this effect by the imbalance and sparsity of training data (Fig. 1), as the stratified sampling strategy for selecting the folds during cross-validation often leads to incomplete training folds (missing rare classes).

In addition to approaches used for run submission, we also experimented with a hierarchical classifier following the *nested dichotomies* approach [2], which trains binary classifiers arranged in a binary tree by recursively dividing the set of classes (and corresponding training samples) into two subsets. However, cross-validation classification performance on the development set was so poor that we excluded this classifier from further experiments. We attribute this failure to the sparsity of training samples, leading to underfitting of several binary classifiers in the tree.

In conclusion, this paper has provided a compelling example of the usefulness of traditional machine learning techniques when combined with CNN-based feature extraction methods for predicting a predefined set of diseases and findings in endoscopic images of the GI tract of the human body. Experiments revealed that both linear support-vector machines and multinomial logistic regression are able to deliver good classification performance at a low run-time complexity in a high-dimensional feature space, learning efficiently from an imbalanced and sparse training set. A more detailed analysis of our approach and a comparison to other submissions will be possible when ground truth labels of the test set and results of all participants have been published.

REFERENCES

[1] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* 12, 6 (2017), e0177678.

[2] Eibe Frank and Stefan Kramer. 2004. Ensembles of Nested Dichotomies for Multi-class Problems. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML '04)*. ACM, 39–46. https://doi.org/10.1145/1015330.1015363

[3] Aurélien Géron. 2017. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc.

[4] Mathias Lux and Savvas A Chatzichristofis. 2008. Lire: Lucene image retrieval: an extensible Java CBIR library. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 1085–1088.

[5] Stefan Petscharnig and Klaus Schöffmann. 2018. Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools and Applications* 77, 7 (2018), 8061–8079. https://doi.org/10.1007/s11042-017-4699-5

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

[7] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 170–174.

[8] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 164–169.

[9] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas de Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *MediaEval 2018 Working Notes*.

[10] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249. https://doi.org/10.1002/widm.1249

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proc. of the IEEE conference on computer vision and pattern recognition*. 1–9.

[12] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann.