Ensemble of Texture Features for finding abnormalities in the Gastro-Intestinal Tract

Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, Muhammad Atif Tahir
School of Computer Science
National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
{k142268,k142092,k142009,atif.tahir}@nu.edu.pk

ABSTRACT

An endoscopy is a procedure in which a doctor uses specialized instruments to view and operate on the internal organs and vessels of the body. This paper aims to predict the diseases and abnormalities in the Gastro-Intestinal Tract, using multimedia data. It differs from other projects in the medical domain because it does not use medical imaging like X-rays, CT scan etc. The dataset, which comprises of 4000 images, is provided by MediaEval Benchmarking Initiative for Multimedia Evaluation. The data is collected during traditional colonoscopy procedures. Techniques from the fields of multimedia content analysis (to extract information from the visual data) and machine learning (for classification) have been used. On testing data, 94% accuracy and an MCC of 0.73 is achieved using logistic regression and ensemble on different features.

1 INTRODUCTION

Medical image diagnosis is one of the most challenging tasks pertinent to the industry of computer vision. Most of the work in the recent times has been done on CT-Scans, X-Rays, and MRI etc. The Medico Task of 2017 challenged their participants to predict the abnormalities in the Gastro-Intestinal tract through endoscopic examination [1]. This implies the presence of multimedia images instead of traditional medical images for the challenge [2]. Deep analysis on GI tract images can help to predict abnormalities and diseases in its initial stages [1]. 4000 images were used for training purpose and the same numbers were reserved for testing data. Different pre-processing techniques were applied and machine learning models were deployed to produce healthy results.

2 OUR PROPOSED APPROACH

Feature Engineering is one of the most challenging and key parts of any Machine Learning Project. Discriminating features are the requirement for function approximation. The task organizers provided 6 pre-computed visual features for every image. These include JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and PHOG.

Since texture plays an important role in the recognition of any object in the image and has been used a lot for different computer vision tasks such as Facial recognition etc. We, therefore, compute the texture of the images using the most common methods of Local Binary Pattern [3] and Haralick features [4]. This drastically improves the classifier accuracy. Through 10-Fold cross validation approach, it was found that some features perform very poorly as compared to others. Hence, they were removed from the model. The refined features were JCD, Edge Histograms, Color Layout, Auto Color Correlogram, Local Binary Pattern with radius 1 and haralick texture features.

We then train separate model using logistic regression [7] and kernel discriminant analysis using spectral regression [5, 6] for each feature because of the composite nature of features. Ensemble technique was then applied to the predictions. Ensemble implies the fact that final model makes use of majority voting among all the independent models trained on each feature. It should be noted that we investigated various advanced machine learning techniques but the best results were obtained using logistic regression and thus reported in this paper.

One of the interesting characteristics of this competition included the limited use of data to train the models. We, therefore, use K-means clustering [8] to come up with a reduced data set representing the whole distribution. We divide the dataset into 10 clusters and extract images from each cluster in an equal ratio. Through this, we extract 732 images from 4000 to train models.

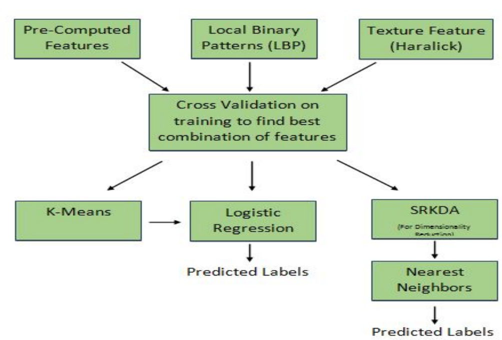

Figure 1: Our proposed model. Pre-computed features are provided by the organizer including ColorLayout, JCD, Edge-Histogram etc.

3 RESULTS AND ANALYSIS

The linear regression model was implemented using Python's scikit-learn package. Among other parameters of logistic regression, two of the most important parameters include "solver" and "multi_class" parameters, for which we used the values of "lbfgs" and "ovr" respectively. The Broyden Fletcher Goldfarb Shanno (BFGS) algorithm is an iterative method for solving unconstrained nonlinear optimization problems. One-Versus-Rest (ovr), also known as one-vs-all, is a strategy which fits one classifier per class. For each classifier, the

MediaEval'17, 13-15 September 2017, Dublin, Ireland

Table 1: Confusion Matrix of the best run (Run_1). n-cecum = normal-cecum. n-z-line = normal-z-line. n-pylorus = normal-pylorus. d= dyed.

Predicted / Actual class	polyps	n-cecum	n-z-line	n-pylorus	esophagitis	d-res-margins	d-lifted-polyps	ulcerative-colitis
polyps	341	13	0	0	0	4	20	98
normal-cecum	71	485	0	0	0	3	4	90
normal-z-line	0	0	451	0	225	0	0	0
normal-pylorus	4	0	1	500	0	0	0	10
esophagitis	3	0	48	0	275	0	0	4
dyed-resection-margins	0	0	0	0	0	360	119	0
dyed-lifted-polyps	5	2	0	0	0	133	356	0
ulcerative-colitis	76	0	0	0	0	0	1	298

class is fitted against all the other classes. In addition to its computational efficiency (only 8 classifiers are needed), one advantage of this approach is its interpretability. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the class by inspecting its corresponding classifier. This is the most commonly used strategy for multiclass classification and is a fair default choice.

We train logistic regression on each feature resulting in 6 different models. Each model provided 8 probabilities, where each probability represented a class confidence score. These probabilities were added together and the class with the highest probability score is chosen to be the predicted label. By applying the proposed model, we obtained the accuracy of 90% with the F1-score of 0.89 and MCC of 0.8 on the training data. While on testing data, which are independently run the organizers, we found the accuracy of 94% with the F-score of 0.76 and MCC of 0.73 (Table 2). The best run is obtained using Run_1 in which all 4000 images are used and this approach is basically ensemble of 6 features (JCD, Edge Histograms, Color Layout, Auto Color Correlogram, Local Binary Pattern with radius 1 and haralick texture features). Logistic regression is being used as the classifier. In summary, following are the 5 runs submitted for the abnormality detection:

3.1 Run_1

Ensemble of 6 features [JCD, Edge Histogram, Color Layout, Auto Color Correlogram, LBP, Haralick] trained on 4000 images, using Logistic Regression.

3.2 Run_2

Same as Run1 but 2000 images were randomly selected.

3.3 Run_3

Same as Run1 with the addition of another feature. The new feature was formulated by Kernel Discriminant Analysis (for dimensionality reduction) which takes an input all the 6 features. For this run, 4000 images were used.

3.4 Run_4

The model was trained on just reduced dimensions which were obtained by KDA. Nearest Neighbour was used as the classifier. Complete training data (4000) is used.

Table 2: Results from testing data independently evaluated by the organizers.

Approach	Precision	Specifity	MCC	F1	Accuracy
Run_1	**0.7665**	**0.966**	**0.736**	**0.767**	**0.942**
Run_2	0.764	0.966	0.734	0.765	0.941
Run_3	0.745	0.963	0.712	0.745	0.936
Run_4	0.564	0.937	0.565	0.509	0.891
Run_5	0.688	0.955	0.649	0.689	0.922

3.5 Run_5

Firstly, KMeans Clustering [8] is applied to obtain 10 clusters from each class. From these clusters, 732 images were selected such that uniformity among the dataset is maintained. Run1 was duplicated on these selected 732 images.

Table 1 shows the confusion matrix of the best run. It is observed that the model performs remarkably well for Normal-Pylorus (all 500 True Positive) and Normal-cecum (485). It also classifies Normal-z-line quite accurately (451), however, Esophagitis is also being confused with Normal-z-line quite often. Polyps are also being correctly classified moderately well (341), however, they are also being confused with Ulcerative-colitis (and vice versa) and Normal-cecum. Lastly, Dyed-resection-margins and Dyed-lifted-polyps are being confused with each other in some cases. It feels like the model is somewhat overfit on the Normal-cecum class.

4 CONCLUSION

We present our proposed model to classify gastro-intestinal abnormalities using endoscopic images. Training (4000 samples) and Testing (4000 samples) data was provided by MediaEval Benchmarking Initiative for Multimedia Evaluation. As mentioned earlier in the introduction, the study used multimedia content analysis, machine learning and ensemble learning techniques for classification. The best of the results were found on logistic regression using ensemble method on 6 different features (including Local Binary Pattern, Haralick texture feature) which resulted in an accuracy of 94% with F1-score of 0.76 and MCC of 0.73 on testing data.

REFERENCES

[1] Michael Riegler, Konstantin Pogorelov, PÃěl Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Losada Eskeland, Duc-Tien Dang-Nguyen, Mathias Lux, Concetto Spampinato, "Multimedia for Medicine: The Medico Task at MediaEval 2017", MediaEvalâĂŹ17, 13-15 September 2017, Dublin, Ireland.

[2] 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, PÃěl Halvorsen, "Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection", Proceedings of ACM on Multimedia Systems Conference (MMSYS), pp.164-169, 2017.

[3] T. Ojala, M. PietikÃďinen, and T. T. MÃďenpÃďÃď, âĂIJMultiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern,âĂİ IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, 2002

[4] Haralick, Robert M., and Karthikeyan Shanmugam. "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics 6 (1973): 610-621.

[5] D. Cai, X. He, and J. Han. Speed up kernel discriminant analysis. The VLDB Journal, 20(1):21âĂŞ33, 2011.

[6] M. A. Tahir et al. A robust and scalable visual category and action recognition system using kernel discriminant analysis with spectral regression. IEEE Transactions on Multimedia, 15(7), 2013.

[7] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

[8] MacQueen, James , Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , 1967