Multimodal Biomedical Image Classification and Retrieval with Multi Response Linear Regression (MLR)-Based Meta Learning

Md Mahmudur Rahman, Prabir Bhattacharya
Computer Science Department,
Morgan State University,
Baltimore, MD, USA
rahmanmm@mail.nih.gov

ABSTRACT

This paper presents a classification-driven biomedical image retrieval approach by combining multiple visual and text features with a multi-response linear regression (MLR)-based meta-learner. Feature descriptors at different levels of image representation are often in diverse forms and complementary in nature. For modality detection of medical images, the MLR has been proposed as a trainable combiner for fusing class probability outputs of several base-level SVM classifiers on different visual and text features as inputs. The advantage of using MLR here over other generalizers is its interpretability as the weights generated by it indicate the different contributions that each features makes for class prediction. Hence, a query-specific adaptive similarity fusion approach is also proposed for image retrieval. Based on the on-line prediction of the query image modalities, individual feature weights generated by MLR are used in a linear combination of similarity matching function for image retrieval. The classification and retrieval results were evaluated evaluated on a standard ImageCLEFmed'2010 benchmark data set of 77,000 images with associated XML annotations and it showed improved performances.

Keywords

medical imaging; multimodality; text retrieval, content-based image retrieval; classification; filtering; data fusion

1. INTRODUCTION

Images are frequently used in biomedical articles to convey essential information or to highlight special cases in context with correlated text [1]. Overall, biomedical literature incorporates an approximation of 100 million figures, whereas the biomedical open access literature of PubMed Central [1] of National Library of Medicine (NLM) alone contained almost four million images in 2016. Retrieval systems, such as the Goldminer [2] search engine and Yale Image Finder (YIF) [2] searching for images within a collection of biomedical articles commonly represent and retrieve them according to their collateral text, such as captions.

Until now, little attention was devoted to the use of image contents in the articles. However, integration of complementary textual and visual information into a unified information retrieval system appears to be promising and could improve retrieval quality through greater utilization of all available (and relevant) information. The results from medical retrieval tracks of previous ImageCLEF campaigns also suggest that the combination of content-based and text-based image searches provides better results than using the two different approaches individually [3]. In order to enable effective search in medical journal articles as well, it might be advantageous for a retrieval system to be able to first recognize the image type (e.g., X-ray, MRI, Ultrasound, etc.). A successful categorization of images would greatly enhance the performance of the retrieval system by filtering out irrelevant images, thereby reducing the search space. In addition, the classification information could be utilized to adjust the weights of different image features (such as; color feature could receive more weight for microscopic and photographic images, and edge- or texture-related features for

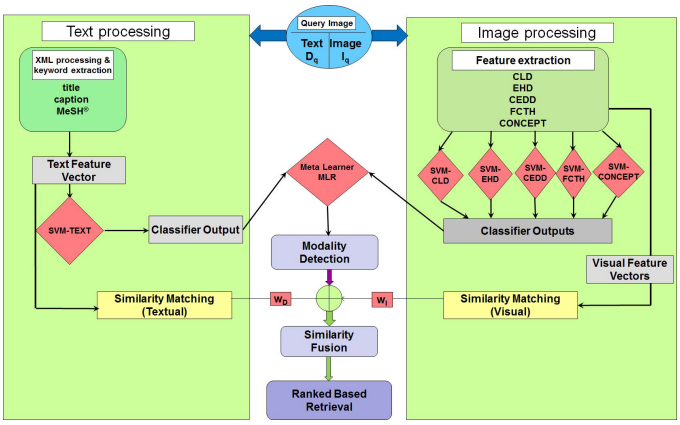

Figure 1: Process flow diagram of the multimodal classification and retrieval approach

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] http://goldminer.arrs.org/

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

radiographs) in similarity matching for a query and database images.

To address a few of the issues described above, and motivated by the successful use of machine learning and fusion techniques in IR, we propose a classification-driven retrieval approach of biomedical images from collections of full-text journal articles. The proposed approach uses text and image features extracted from relevant components in a document, and utilizes the multi response linear regression (MLR) [5] as meta combiner for both classifier combination and similarity fusion. The advantage of using MLR over other generalizers is its interpretability as it provides a method of combining the confidence generated by the base level models into a final decision. The weights generated by MLR indicate the different contributions that each base level model makes to the prediction classes. The data flow diagram of the multimodal retrieval process is shown in Fig. 1.

2. MODALITY DETECTION WITH MLR

The modality detection is an important task toward achieving high performance in biomedical image retrieval. The variation of the medical image categories at this global level can be effectively modeled by using any supervised learning techniques. Owing to their empirical success, we used multi-class SVMs [6] as base-level classifiers for classifying images into different modalities based on their textual and visual features. The classification is performed by combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC) [9]. For the SVM training, the input is a feature vector set of training images and a set of m labels are defined as $\{\omega_1, \cdots, \omega_i, \cdots, \omega_m\}$, where each ω_i characterizes the representative image modality. In this context, given a feature vector $\mathbf{x}$, the multi-class SVM estimates the probability or confidence scores of each category as

$$p_i = P(y = \omega_i \mid \mathbf{x}), \text{for } 1 \leq \mathrm{i} \leq \mathrm{m} \tag{1}$$

The feature descriptors are often in diversified forms and complementary in nature. Hence, multiple classifiers are needed to deal with different features resulting in a general problem of combining those classifiers to yield improved performance. The combination of ensembles of classifiers has been studied intensively and evaluated on various image classification data sets involving the classification of digits, faces, photographs, etc. [10]. In this work, we used MLR as a trainable combiner for combining outputs of base-level SVM classifiers which are trained with individual features as inputs. MLR attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data [5]. In recent years, MLR has been recommended as a combiner for merging heterogeneous base-level classifiers [11, 12]. Any classification problem with real-valued attributes can be transformed into a multi-response regression problem.

In [12], it is found that to handle multi-class problems, the best results are obtained when the higher-level model combines the confidence (and not just the predictions) of the lower-level ones based on the MLR algorithm. Based on this finding, we used MLR where the variables are confidence scores (probabilistic outputs) based on the base-level SVM classifier's prediction on each feature for each modality class based on Equation 1. Given a data set $\mathcal{L} = \{(y_n, x_n), n = 1, \cdots, N\}$, where y_n is a class label taking values from one of the m classes and $\mathbf{x}_n$ is a vector representing the attribute values of a feature (instance) x_n of the $n-th$ instance. Now, for K different features, K base-level classifiers $C_1, C_2, \cdots, C_K$ are generated by applying a multi-class SVM learning algorithm [9].

So, each base-level classifier is trained with a particular input feature x_n from the training set. The prediction of classifier $C_k (k = 1, 2, \cdots, K)$ when applied to a feature vector $\mathbf{x}_n$ is a probability distribution vector

$$\begin{aligned} \mathbf{P}_k(\mathbf{x}_n) = & \quad [P_k(\omega_1|\mathbf{x}_n), P_k(\omega_2|\mathbf{x}_n), \cdots, P_k(\omega_m|\mathbf{x}_n)]^{\mathrm{T}} \\ \doteq & \quad [P_k^1(x_n), P_k^2(x_n), \cdots, P_k^m(x_n)]^{\mathrm{T}}, k = 1, 2, \cdots, K. \end{aligned} \tag{2}$$

where $P_k^m(x_n)$ denotes the probability or class confidence score that the feature vector $\mathbf{x}_n$ of example x_n belongs to class ω_m as estimated by the classifier C_k. Furthermore, $\mathbf{P}(\mathbf{x}_n)$ is defined as an mK-dimensional column vector as

$$\begin{aligned} & \mathbf{P}(\mathbf{x}_n) = [\mathbf{P}_1(\mathbf{x}_n), \cdots, \mathbf{P}_k(\mathbf{x}_n), \cdots, \mathbf{P}_K(\mathbf{x}_n)]^{\mathrm{T}} \\ & = [P_1^1(x_n), P_1^2(x_n), \cdots, P_1^m(x_n), P_2^1(x_n), P_2^2(x_n), \\ & \cdots, P_2^m(x_n), \cdots, P_K^1(x_n), P_K^2(x_n), \cdots, P_K^m(x_n)]^{\mathrm{T}} \end{aligned} \tag{3}$$

Based on the intermediate feature space constituted by the outputs of each base-level SVM classifier, the MLR method [11] firstly transforms the original classification task with m classes into m regression problems: the problem for class ω_j has examples with responses equal to one when they indeed have class label ω_j and zero otherwise.

For each class ω_j, MLR selects only $P_1^j(x_n), P_2^j(x_n), \cdots, P_K^j(x_n)$, the probabilities that x_n belongs to ω_j predicted by the base-level classifiers $C_1, C_2, \cdots, C_K$, as the input attributes to establish a linear equation

$$LR_j(x_n) = \sum_{k=1}^{K} \alpha_k^j \; P_k^j(x_n), j = 1, 2, \cdots, m \tag{4}$$

where the coefficients $\{\alpha_k^j\}$ are constrained to be non-negative and the nonnegative-coefficient least-squares algorithm described in [5] is employed to estimate them.

To classify a new instance x, we need to compute $LR_j(x)$ for all the m classes and assign it to the class ω_j which has the greatest value:

$$LR_j(x) > LR_{j'}(x) \;\; \text{for all } \mathrm{j}' \neq \mathrm{j}. \tag{5}$$

The advantage of using MLR over other generalizers is its interpretability as it provides a method of combining the confidence (class probabilities) generated by the base level models into a final decision. The weights generated by MLR indicate the different contributions that each base level model (e.g., features) makes to the prediction classes.

3. CLASSIFICATION-DRIVEN SIMILARITY FUSION

One of the most commonly used approaches of similarity fusion in image retrieval is the linear combination of similarity scores of different features with pre-determined weights. In this approach, the similarity between a query image I_q and target image I_j is described as

$$\mathrm{Sim}(I_q, I_j) = \sum_F \alpha^F \; \mathrm{S}^F(I_q, I_j) = \sum_F \alpha^F \; \mathrm{S}(\mathbf{f}_q^{\mathrm{F}}, \mathbf{f}_j^{\mathrm{F}}) \tag{6}$$

where $F \in \{\text{Color, Texture, Edge, Shape, Keyword, etc.}\}$ and $S(\mathbf{f}_q^F, \mathbf{f}_j^F)$ are the similarity matching function in individual feature spaces and α^F be the weights generally hard coded in the systems within the different feature representation schemes where a color feature will have the same weight for the search of both microscopic pathology or X-ray images.

We proposed a dynamic linear combination scheme where based on the on-line category prediction of a query image, pre-computed category-specific feature weights (e.g., α^F) in Equation 4 as generated by MLR, are utilized in the linear combination of the similarity matching function as depicted in Algorithm 1.

Algorithm 1 Category-Specific Similarity Fusion with MLR generated weights

(Off-line): Store category specific feature-weights from the learning of MLR combiner.
(On-line): For a query image I_q, calculate individual feature vectors $\mathbf{f}_q^F$.
For each feature, get a category prediction based on the probabilistic output of Equation ((1)).
Classify image I_q as $\omega_i(q), i \in \{1, \cdots, m\}$ based on Formula (5) of MLR combiner.
Consider the individual features weights α^F generated by Equation (4) for the query image category $\omega_i(q)$.
Combine the similarity scores with the weights based on similarity fusion in Formula (6).
Finally, return the images based on the similarity matching values in descending order to obtain a final ranked list of images.

4. EXPERIMENTS AND RESULTS

To evaluate the retrieval effectiveness, experiments are performed on the ImageCLEFmed'2010 [4] benchmark medical image collection of nearly 77,500 images from over 5,600 articles. For classification experiment, we used a collection of around 9500 images where images are classified into one of the 8 modalities (e.g., CT, MR, XR, etc,) Many of these images were collected from training sets of modality detection task of ImageCLEFmed in the last several years. The data set is divided with a training set of a 2100 images and the rest of the images are used for testing with almost similar distribution of image categories. The experimental results are generated based on the 16 ad-hoc query topics where each topic is consisted of the text keywords itself in three languages (English, German, French) and 2 to 3 example images for the visual part of the topic.

Each image in the collection is annotated as a XML document of image-related text. In our case, information from the title and caption tags are extracted and preprocessed by removing stop words. Subsequently, the remaining words are reduced to their stems, which finally form the index terms or keywords. The images are presented as a vector of words based on the popular vector space model (VSM) of Information Retrieval. For visual feature, MPEG-7 based Color Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD) and descriptors from the Lucene Image REtrieval (LIRE) library [8], such as Fuzzy Color Texture Histogram (FCTH) and Color Edge Direction Descriptor (CEDD) are extracted to represent images from different perspectives. In addition, a "Bag of Visual Words" based feature is used by mapping local patches to "concepts" using supervised classification technique [7].

Table 1: Test Results (7465 Images)

Feature	Accuracy
CLD	64.43%
EHD	65.33%
CEDD	73.95%
FCTH	70.60%
Concept	76.93%
Text	70.83%
Combined (Visual)	80.85%
Combined (Visual+text)	85.30%

Table 2: MLR weights based on individual visual and text features

Class	CLD	EHD	CEDD	FCTH	Concept	Text
CT	0.05	0.26	0.20	0.14	0.28	0.74
MC	0.02	0.32	0.24	0.11	0.31	0.38
MR	0.11	0.02	0.07	0.02	0.32	0.79
PT	0.04	0.05	0.06	0.08	0.06	0.28
PH	0.06	0.07	0.20	0.22	0.11	0.16
US	0.32	0.31	0.16	0.25	0.59	0.27
XR	0.09	0.24	0.08	0.16	0.44	0.43
GX	0.02	0.24	0.18	0.10	0.48	0.45

The results of the modality classification approaches were compared using classification accuracy. For the base-level SVM classifier learning on different image features, the radial basis function (RBF) as kernel is used. A 10-fold cross-validation (CV) is conducted for each feature to find the best values of the tunable parameters C and γ of the RBF kernel, which are utilized for the final training to generate the SVM model files. Table 1 shows the classification results on the test data set of 7465 images. The best accuracy (85.30%) is achieved when classification is performed in the combined feature space (visual+text), but at the computational expense of a much larger feature vector.

In order to estimate the feature weights (e.g., coefficient $\{\alpha_k^j\}$ in formula (4)), or more generally to train the combiner MLR, we have to form the meta-level data. Based on the findings in [12], we used MLR as a trainable combiner by employing the validation strategy. Hence, the training set (2100 images) is used to derive the base-level SVM classifiers as mentioned above and the test set (7465 images) is employed to construct the meta-level data based on the probabilistic outputs generated by the SVM classifiers on the test images. For example, Table 2, shows the weights generated by meta learner MLR based on using individual image and text features. These weights are finally used for classifier combination to classify an image by using the formula (4) and formula (5).

Table 3 shows the final classification results on the test data set based on using either equal or MLR weights on different feature space combinations. In all three cases, we observe that MLR combiner improves accuracy around 2-3% compared to using equal feature weighting and the best

Table 3: Classification Accuracies with Equal and MLR weights (Test Set)

Feature	Equal Weight	MLR Weight
Individual Visual Only	82.76%	84.91%
Individual Visual + Text	86.48%	90.12%
Combined Visual + Text	86.71%	87.76%

Table 4: Retrieval Results based on the 16 Ad-Hoc Topics

Feature	MAP	GMAP	B-Pref
Indv. Visual (Equal)	0.0024	0.0002	0.0150
Indv. Visual (MLR)	0.0025	0.0002	0.0172
Text	0.1195	0.0095	0.1566
Indv. Visual + Text (Equal)	0.0730	0.0073	0.1185
Indv. Visual + Text (MLR)	0.1240	0.0167	0.1608
Comb. Visual + Text (Equal)	0.1260	0.0213	0.1655
Comb. Visual + Text (MLR)	0.1310	0.0201	0.1677

accuracy (90.12%) is achieved when each visual and text features were weighted separately by MLR. As expected, combining classifiers with category-specific individual feature weights on complementary features benefits the performance.

Table 4 shows the retrieval result in the ImageCLEFmed'2010 benchmark medical image collection. It is observed that the best MAP score (0.13) is achieved when a multi-modal search is performed in a MLR-weighted combined visual and text feature spaces. We can also observe a large improvement in MAP score (0.1240) when individual visual features are weighted based on the image category (e.g. MLR weighted) instead of using equal weighting approach. The other scores (e.g., GMAP, Rprec, and Bpref) also slightly improved in most cases with MLR-based weighting approach compared to equal weighting. To breakdown the results further, Fig. 2 shows the bar graph (chart) of multi-modal searches based on precision over the top K (5, 10, and 20) images. We observed significant improvement of precision at these early levels for the multi-modal searches with MLR-based weighting by using both individual and combined visual feature spaces.

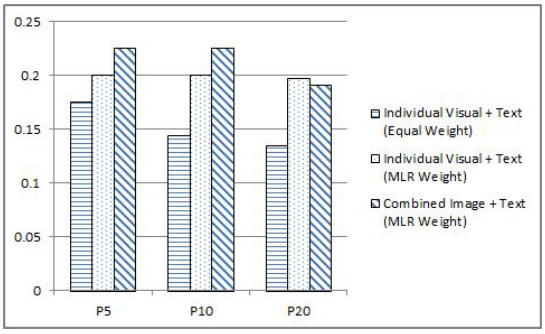

Figure 2: Precision at P5,P10, and P20 for different search modalities

5. CONCLUSIONS

A novel multi-modal retrieval approach is proposed by utilizing the classification result of a MLR meta-learner directly in the retrieval loop. The benchmark ImageCLEFmed'2010 collection with a query set and corresponding performance measure model provided enough reliability for objective performance evaluation. Our retrieval results demonstrate the effectiveness of the search approaches compared to using only a single modality or without using any classification information.

6. REFERENCES

[1] Demner-Fushman, D., Antani, S. K., Simpson, M., and Thoma, G. R. 2009. Annotation and retrieval of clinically relevant images. *Int J Med Inform.*, 78 (12), (Dec. 2009), e59–67. DOI= 10.1016/j.ijmedinf.2009.05.003.

[2] Xu, S., McCusker, J., and Krauthammer, M. 2008. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics.* 24(17) (2008), 1968–70.

[3] Kalpathy-Cramer, J., de Herrera A.G., Demner-Fushman, D., Antani, S.K., Bedrick, S. and MÃijller, H. (2015). Evaluating performance of biomedical image retrieval systems–an overview of the medical image retrieval task at ImageCLEF 2004-2013. Comput Med Imaging Graph. 39, 55-61.

[4] Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Reisetter, Jr. J., C.E.K., Hersh, W. R. 2010. Overview of the CLEF 2010 Medical Image Retrieval Track, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes.* Padua, Italy, (Sep. 2010)

[5] Lawson, C. J., Hanson, R. J. 1995. Solving Least Squares Problems. SIAM Publications, Philadephia.

[6] Vapnik, V. 1998. Statistical Learning Theory. New York, NY, Wiley.

[7] Rahman, M. M., Antani, S. K. and Thoma, G. R. 2009. A Medical Image Retrieval Framework in Correlation Enhanced Visual Concept Feature Space. *Proc. 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, August 3-4, 2009, Albuquerque, New Mexico, USA.

[8] Mathias L. and Savvas A. C. 2008. Lire: lucene image retrieval: an extensible java CBIR library. *Proceedings of the 16th ACM international conference on Multimedia*, October 26-31, 2008, Vancouver, British Columbia, Canada, 1085-1088.

[9] Wu, T. F., Lin, C. J. and Weng, R. C. 2004. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. of Mach. Learn. Research.* 5 (2004), 975–1005.

[10] Kittler, J., Hatef, M., Duin, R. P. W. and Matas, J. 1998. On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 20 (3) (1998), 226–239.

[11] Chun-Xia, Z. and Robert P. W. 2009. An Empirical Study of a Linear Regression Combiner on Multi-class Data Sets. *LNCS*, 5519 (2009), 478âĂŞ-487. Springer-Verlag Berlin Heidelberg 2009

[12] Ting, K. M. and Witten, I. H. 1999. Issues in Stacked Generalization. *Journal Of Artificial Intelligence Research.* 10 (1999), 271–289.