Classification with reject option using contextual information

Filipe Condessa
Universidade de Lisboa

Jose Bioucas-Dias
Universidade de Lisboa

Carlos A. Castro
University of Pittsburgh

John A. Ozolek
Children’s Hospital of Pittsburgh

Jelena Kovacevic
Carnegie Mellon University, jelenak@cmu.edu

Follow this and additional works at: http://repository.cmu.edu/ece

Part of the Biomedical Engineering and Bioengineering Commons, and the Electrical and Computer Engineering Commons

Published In
CLASSIFICATION WITH REJECT OPTION USING CONTEXTUAL INFORMATION

Filipe Condessa¹²,José Bioucas-Dias¹², Carlos A. Castro⁵, John A. Ozolek⁶, and Jelena Kovačević³⁴

¹Inst. de Telecomunicações & ²Dept. of ECE Inst. Superior Técnico Technical University of Lisbon, Lisbon, Portugal
³Dept. of BME and Center for Bioimage Informatics, ⁴Dept. of ECE Carnegie Mellon University, Pittsburgh, PA, USA
⁵Dept. of Obstetrics and Gynecology, Magee-Womens Research Inst., Foundation Univ. of Pittsburgh
⁶Dept. of Pathology, Children’s Hospital of Pittsburgh University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

ABSTRACT

We propose a new algorithm for classification that merges classification with reject option with classification using contextual information. A reject option is desired in many image-classification applications requiring a robust classifier and when the need for high classification accuracy surpasses the need to classify the entire image. Moreover, our algorithm improves the classifier performance by including local and nonlocal contextual information, at the expense of rejecting a fraction of the samples. As a probabilistic model, we adopt a multinomial logistic regression. We use a discriminative random model for the description of the problem; we introduce reject option into the classification problem through association potential, and contextual information through interaction potential. We validate the method on the images of H&E-stained teratoma tissues and show the increase in the classifier performance when rejecting part of the assigned class labels.

Index Terms—image classification, reject option, discriminative random fields

1. INTRODUCTION

Classification is a ubiquitous image processing task that aims to separate a group of objects into different classes. In classification, small, unbalanced, or incomplete training sets can lead to low performance of the classifier. As in many applications, the need for high accuracy surpasses the need to classify all the samples, in those applications, we classify while rejecting a portion of the class labels [1]. To aid in the process, one can exploit the similarity between samples and the spatial context, providing useful cues when classifying. Classifying with rejection as well as using spatial contextual information are both applicable in the automated identification of tissues in histopathology [2, 3, 4], where the cost of creating a large and representative training set is high, the presence of unknown classes is a possibility, and the similarity between tissues belonging to the same class is high.

Our goal is thus to improve classifier performance by adding a reject option and contextual information as follows:

• Partition the image; perform feature extraction in each partition.
• Estimate class probabilities for each partition and the risk associated with each class.
• Classify each partition using contextual information and the reject option.

The paper is organized as follows: Section 2 provides the background on partitioning, feature extraction and classification techniques. Section 3 introduces the concept of rejection while Section 4 describes our classification method with reject option using contextual information. Section 5 presents experimental results and Section 6 concludes the paper.

2. BACKGROUND

We now briefly describe the background necessary for our work in terms of image partitioning, feature extraction procedures, and the classification methods used. A generic classifier consists of a feature extraction function that maps a high-dimensional input space into a lower-dimensional feature space, and a classification function that maps the feature space into the class label space. The partitioning of the image can be considered as a preprocessing method, mapping the image space into the partition space.

Partitioning To reduce the dimensionality of the problem, we partition the data, which also allows us to efficiently use graph-based methods. The partitioning of the image is performed by oversegmenting the image and creating superpixels [5]; this allows us not only to reduce the dimension of a problem by a factor of the order of 1000, but also guarantees that there is class and appearance coherence in each partition, due to the small dimension of each superpixel (each superpixel is of average size $5 \times 10^2$ pixels, corresponding to $4 \times 10^3$ partitions in a $1600 \times 1200$ image). One drawback of this partitioning method is the nonuniformity of the partitions in shape and size.

Feature Extraction We use two different types of features in our work: application-specific features and similarity features. We normalize features from each partition by its mean and standard deviation. As application-specific features we use the histopathology vocabulary (HV) containing features with physiological relevance,
Classification

Given the set of partitions $S$ and associated sets of features, we want to classify each partition into a single class. As we need probability estimates for that task, we use multinomial logistic regression (MLR) [6]. A linear combination of kernels is used as the nonlinear regression function to include both the application-specific as well as the similarity features.

Multinomial Logistic Regression. We model the a posteriori probability $p(y_i \mid f_i, W)$ with MLR, where $f_i$ is the feature vector of the $i$th partition, $\mathcal{L} = \{1, 2, \ldots, N\}$ set of class labels that can be assigned to each partition, $y_i \in \mathcal{L}$ the class label assigned to the $i$th partition, and $W$ the regression matrix with columns $w_i$, $i = 1, 2, \ldots, N$ (since $W$ is translation invariant, we arbitrarily set $w_N = 0$). Let $T$ be the set of indices of the partitions present in the training set and $\mathcal{F}_T = \{f_i\}_{i \in T}$. Then, the MLR models a posteriori probabilities as

$$p(y_i = \ell \mid f_i, W_T, W) = \frac{e^{w_T^T k(f_i, f_T)}}{\sum_{j=1}^{N} e^{w_T^T k(f_i, f_T)}},$$

where $k(f_i, f_T)$ is a kernel vector obtained by concatenating two length-|$T$| vectors of application-specific features $k_i(f_i, f_T)$ and similarity features $k_s(f_i, f_T)$.

LORSAL. To avoid overfitting and ensure generalization capability of the classifier, we adopt an element-wise independent Laplacian prior for the MLR vector and compute the maximum a posterior estimate of $W$ by solving the optimization problem

$$\hat{W} = \arg \max_W \{l(W) + \log p(W)\},$$

with $l(W) = \sum_{i \in S} \log p(y_i \mid f_i, W)$, $p(W) = \alpha e^{-\lambda_1 W_1}$, where $\lambda$ is the regularization parameter. Note that $\log p(W)$ is proportional to the negative of the $\ell_1$ norm of $W$, which promotes sparse regression matrices, controlled by $\lambda$. The maximization (1) is performed with the Logistic Regression via Splitting and Augmented Lagrangian (LORSAL) algorithm [7], which solves the equivalent problem

$$\arg \max_W \{l(W) + \log p(\Omega)\}, \quad \text{subject to } W = \Omega.$$

3. REJECT OPTION

To improve accuracy at the expense of not classifying all the partitions, we classify while rejecting. Let $\mathcal{L}' = \mathcal{L} \cup \{N + 1\}$ be an extended set of partition class labels with an extra label. We consider two different class rejection concepts: uninteresting class, present in the training set, trained as a regular class and uninteresting to the observer (corresponding to the class label $y_i = 1$), and unknown class, which arises from an inability of the classifier to correctly classify all labels (corresponding to the class label $y_i = N + 1$).

Let $\mathcal{C}$ and $\mathcal{Z}$ be the sets of labels of correctly and incorrectly classified samples, respectively, and $rc$ and $rz$ be the rejection ratios for correctly and incorrectly classified samples, respectively. Define the accuracy of nonrejected samples as

$$A_{nr} = \frac{(1 - rc) |\mathcal{C}|}{(1 - rz) |\mathcal{Z}| + (1 - rc) |\mathcal{C}|},$$

with the rejection ratio

$$r = \frac{rz |\mathcal{Z}| + rc |\mathcal{C}|}{|\mathcal{Z}| + |\mathcal{C}|}.$$
Association Potential We introduce reject option into the classification problem through association potential. Given the risk function $r(y_i \mid f_i, W)$ from (2), the association potential is

$$V_A(y_i) = \log (r(y_i \mid f_i, W)).$$

Interaction Potential We introduce contextual information into the classification problem through interaction potential [11] of neighborhoods. A local neighborhood $E_{i,n} \subset E$ is the set of edges connecting partition $i$ and its immediate neighbors. A nonlocal neighborhood $E_{i,n} \subset E$ is the set of edges connecting partition $i$ and other partitions with high similarity. This similarity is determined by the similarity function $s : (i, j) \rightarrow [0, 1]$ that assigns a similarity value between two partitions $i$ and $j$. The similarity function is based on features that better assess similarity, which are not being present in the classifier. For example, we can feed the image into a Gabor filter bank and then make a similarity decision based on the Gabor coefficients of the two partitions.

Let $\phi$ be a transition function between two partitions, and $\psi$, $\psi_n$ the weights associated with local and nonlocal neighborhoods, respectively. Then, the interaction potential can be written as a sum of individual (possibly overlapping) interaction potentials,

$$\sum_{(i,j) \in E} V_i(y_i, y_j) = \psi \sum_{(i,j) \in E} \phi(y_i, y_j) + \psi_n \sum_{(i,j) \in E_n} \phi(y_i, y_j).$$

The transition function $\phi$ enforces (1) piecewise smoothness (in a given concept of neighborhood), (2) ease of transition to the unknown class ($y_i = N + 1$), (3) ease of transition between the uninteresting ($y_i = 1$) and unknown ($y_i = N + 1$) classes,

$$\phi(y_i, y_j) \equiv \begin{cases} 1 - \delta_{y_i, y_j}, & y_i, y_j \neq N + 1; \\ 0, & \text{otherwise.} \end{cases}$$

The local neighborhood enforces a local piecewise smoothness of the labeling, while the nonlocal neighborhood enforces a nonlocal piecewise smoothness of the labeling.

5. EXPERIMENTAL RESULTS

For each image in the data set, the method is applied multiple times with randomized training sets. We test the method using training sets based on both a single image as well as the entire data set.

Data Set The data set consists of 36,1600 × 1200 images of H&E-stained teratoma tissues imaged at 40X magnification containing 23 classes. We show results on three representative images in Figure 1, using the entire data set for the training samples in the multiple-image training set approach.

Methods We analyze the performance based on the following criteria: the comparison between the initial accuracy $A_{ini}$ obtained by the labeling resulting from a maximum a posteriori $\hat{p}(f_i, W)$ and the final accuracy $A_{fin}$ obtained by the labeling (4); the improvement ratio $A_{imp} = A_{fin}/A_{ini}$; the rejection rate $r$; and rejection ratios for correctly and incorrectly classified samples $r_C$ and $r_Z$, respectively, and the ratio between the two, $r_Z/r_C$.

For each image $X_i$, we define four training sets: (1) $T_{S1}$ consists of 60 randomly selected partitions of $X_i$ ($\sim 1.5\%$ of $X_i$). (2) $T_{S2}$ consists of 600 randomly selected partitions of $X_i$ ($\sim 15\%$ of $X_i$). Let us consider an auxiliary training set $T_{aux}$ consisting of 30 randomly selected partitions all the images except $X_i$. (3) $T_{M1}$ consists of the union of $T_{S1}$ and $T_{aux}$. (4) $T_{M2}$ consists of the union of $T_{S2}$ and $T_{aux}$.

Results From Table 1, we see that there is a substantial accuracy improvement by including the reject option and contextual information in the classification. Table 2 shows the effect of different rejection ratios on the accuracy of the labeling for a single example image. There is a tradeoff between accuracy improvement, the performance of the rejection option (quantified by $r_Z/r_C$) and the overall rejection rate. Note that $A_{ini}$ in Table 2 is obtained for a single instantiation of $T_{S1}$, and is thus different from that in Table 1, which is obtained as the average of 30 instantiations of $T_{S1}$. Figure 1 shows results on the three example images.

<table>
<thead>
<tr>
<th>Images</th>
<th>$A_{ini}$</th>
<th>$A_{fin}$</th>
<th>$A_{imp}$</th>
<th>$r$</th>
<th>$r_Z$</th>
<th>$r_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set $T_{S1}$</td>
<td>(1) 0.7872</td>
<td>0.8396</td>
<td>1.0671</td>
<td>0.0905</td>
<td>0.2930</td>
<td>0.0388</td>
</tr>
<tr>
<td></td>
<td>(2) 0.6053</td>
<td>0.7272</td>
<td>1.2043</td>
<td>0.2156</td>
<td>0.4178</td>
<td>0.0746</td>
</tr>
<tr>
<td></td>
<td>(3) 0.6109</td>
<td>0.7316</td>
<td>1.1884</td>
<td>0.1967</td>
<td>0.4098</td>
<td>0.0618</td>
</tr>
<tr>
<td>Training set $T_{S2}$</td>
<td>(1) 0.9068</td>
<td>0.9431</td>
<td>1.0401</td>
<td>0.0555</td>
<td>0.3726</td>
<td>0.0247</td>
</tr>
<tr>
<td></td>
<td>(2) 0.7856</td>
<td>0.8885</td>
<td>1.1310</td>
<td>0.1295</td>
<td>0.4913</td>
<td>0.0303</td>
</tr>
<tr>
<td></td>
<td>(3) 0.8256</td>
<td>0.9137</td>
<td>1.1068</td>
<td>0.1086</td>
<td>0.4997</td>
<td>0.0279</td>
</tr>
<tr>
<td>Training set $T_{M1}$</td>
<td>(1) 0.6628</td>
<td>0.8240</td>
<td>1.2448</td>
<td>0.2060</td>
<td>0.5417</td>
<td>0.0407</td>
</tr>
<tr>
<td></td>
<td>(2) 0.4463</td>
<td>0.5888</td>
<td>1.3392</td>
<td>0.3145</td>
<td>0.4701</td>
<td>0.1054</td>
</tr>
<tr>
<td></td>
<td>(3) 0.3620</td>
<td>0.5041</td>
<td>1.4111</td>
<td>0.3561</td>
<td>0.4998</td>
<td>0.1123</td>
</tr>
<tr>
<td>Training set $T_{M2}$</td>
<td>(1) 0.7207</td>
<td>0.8613</td>
<td>1.1940</td>
<td>0.1678</td>
<td>0.5391</td>
<td>0.0341</td>
</tr>
<tr>
<td></td>
<td>(2) 0.7176</td>
<td>0.8337</td>
<td>1.1651</td>
<td>0.1640</td>
<td>0.4727</td>
<td>0.0364</td>
</tr>
<tr>
<td></td>
<td>(3) 0.7296</td>
<td>0.8812</td>
<td>1.2131</td>
<td>0.1853</td>
<td>0.5943</td>
<td>0.0317</td>
</tr>
</tbody>
</table>

Table 1. Average values for the performance metrics for single-image and multiple-image training sets obtained after 30 randomized runs of the classification method corresponding to example images (1), (2) and (3) in Figure 1.

<table>
<thead>
<tr>
<th>$A_{ini}$</th>
<th>$A_{fin}$</th>
<th>$A_{imp}$</th>
<th>$r$</th>
<th>$r_Z$</th>
<th>$r_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>10^5</td>
<td>0.9300</td>
<td>0.5300</td>
<td>0.1300</td>
<td>0.010</td>
</tr>
<tr>
<td>$r$</td>
<td>0.0000</td>
<td>0.0503</td>
<td>0.1009</td>
<td>0.1992</td>
<td>0.5715</td>
</tr>
<tr>
<td>$r_Z$</td>
<td>0.0000</td>
<td>0.1677</td>
<td>0.3468</td>
<td>0.5403</td>
<td>0.9274</td>
</tr>
<tr>
<td>$r_C$</td>
<td>0.0000</td>
<td>0.0234</td>
<td>0.0457</td>
<td>0.1224</td>
<td>0.4915</td>
</tr>
<tr>
<td>$r_Z/r_C$</td>
<td>7.0167</td>
<td>7.5886</td>
<td>4.4142</td>
<td>1.8689</td>
<td>1.5921</td>
</tr>
<tr>
<td>$A_{fin}$</td>
<td>0.8390</td>
<td>0.8582</td>
<td>0.8687</td>
<td>0.8947</td>
<td>0.9689</td>
</tr>
</tbody>
</table>

Table 2. Effect of changing the parameter $\rho$ in (2) for Image (1) in the $T_{S1}$ training set.

6. CONCLUSIONS AND FUTURE RESEARCH

The inclusion of a reject option using spatial contextual information greatly improves the accuracy of the classification. The relative high weight of the rejection rate for incorrect classifications compared to the rejection rate for correct classifications ($r_Z/r_C$) points to a good behavior of the reject option. These encouraging results point towards potential utility in large-scale automated tissue identification of histological slices. Possible future research directions are: (1) Use of a partitioning algorithm that includes more specific features (with physiological relevance). (2) Modeling the transition function $\phi$ using expert knowledge of tissue transitions in normal
and abnormal histological images. (3) Inclusion of multiscale partitioning, classification and rejection in the interaction and association potentials.

7. REFERENCES


