Active Learning with Interactive Response Time and its Application to the Diagnosis of Parasites

Priscila T. M. Saito[†⋆], Pedro J. de Rezende[†], Alexandre X. Falcão[†]
[†] Institute of Computing, University of Campinas, Brazil
[⋆] Department of Computing, Federal University of Technology - Parana, Brazil
Email: {maeda, afalcao, rezende}@ic.unicamp.br

***Abstract*—We have developed an automated system for the diagnosis of intestinal parasites from optical microscopy images. Each exam produces about 2,000 images with hundreds of objects in each image for classification as one out of the 15 most common species of parasites or impurity. As the number of exams increases, a dataset with unlabeled samples for classification grows in size. Impurities are numerous and diverse, with similar features to several species of parasites. Some species are also difficult to be differentiated. In this context, datasets are large and unbalanced, making the identification of the best samples for expert supervision crucial for the design of an effective classifier. We have addressed the problem by proposing a new paradigm for active learning, in which the dataset can be *a priori* reduced and/or organized to make that process realistic (efficient) for user interaction and yet more effective. We have also proposed several active learning methods under this paradigm and evaluated them for the diagnosis of intestinal parasites and other applications. Data reduction and/or organization avoid to reprocess the large dataset at each learning iteration, enabling to halt sample selection after a desired number of samples per iteration, which yields interactive response times. The proposed methods were validated in comparison with state-of-the-art approaches. Experiments included three datasets with parasites and/or impurities. One with 1,944 parasites (without impurities) and another with almost 6,000 labeled objects were used to develop the methods. A more realistic one, with over 140,000 unlabeled objects, unbalanced classes, absence of classes, and considerably higher number of impurities, was used for final validation by an expert in Parasitology.**

***Keywords*-Active learning; pattern recognition; automated diagnosis of intestinal parasites; microscopy image analysis; optimum-path forest classifiers.**

I. Introduction

Nowadays, large datasets are easily found in a wide range of real-world applications, due to the advances in data acquisition and storage. Their fully annotation considerably facilitates information organization, retrieval, and prediction, but it is infeasible for human beings [1]. The task can be accomplished by an effective pattern classifier, but expert interaction is still needed to construct a training set with labeled samples. For the sake of effectiveness in the construction of the training set, the under-developing classifier can assist in the identification of the best samples for expert supervision. Such samples should represent all classes and, in this context, uncertain (informative) samples can provide a fine adjustment by helping the classifier to discriminate among classes with similar properties.

[1]This work relates to a Ph.D. thesis

Active learning methods have been proposed to address the problem [2], [3]. However, for the sake of efficiency, they should reduce as much as possible the response time between user interactions, the training set size, and the number of learning iterations. These aspects seem to not have caught much attention in the literature until the present work.

Traditional techniques for active learning usually follow as general strategy the classification and organization of the entire dataset at every iteration in order to select and present a limited number of labeled samples for expert supervision (Figure 1a). This strategy is not suitable to be applied to large datasets, making the response time not interactive. The present paper is related to the Ph.D. Thesis [4], which addressed the problem by proposing a novel active learning paradigm based on *a priori* data reduction and/or organization, so that the classifier can identify the best samples considerably faster, at interactive times for real applications (Figure 1b). We have also proposed some active learning techniques, evaluated them on several datasets with respect to other approaches, and validated the most suitable one for a real application: the automated diagnosis of human intestinal parasites — a problem that we have investigated for over 10 years.

II. Materials

Image acquisition, segmentation, and feature description were not the focus of this project as they had been previously developed by our group [5], [6]. Our work started from the unlabeled large dataset that results from this process.

For the image database construction, fecal samples were collected from endemic areas of the state of São Paulo: university hospitals at the University of Campinas (UNICAMP) and at the São Paulo State University (UNESP), as well as the Ouro Verde Hospital in Campinas. They were processed at the Visual Computing Laboratory in Biomedical and Health at the Institute of Computing, UNICAMP.

After fecal sample processing, microscope slides were prepared for automatic image acquisition by using a computer-controlled system with microscope, digital camera, focus drive, and motorized stage. Each slide can produce hundreds of objects, obtained by image segmentation, from about 2,000 images of 4M pixels each. Image segmentation was performed by a method based on the image foresting transform [7]. The objects are single components, candidates to be classified as impurity or one out of the 15 most common species of

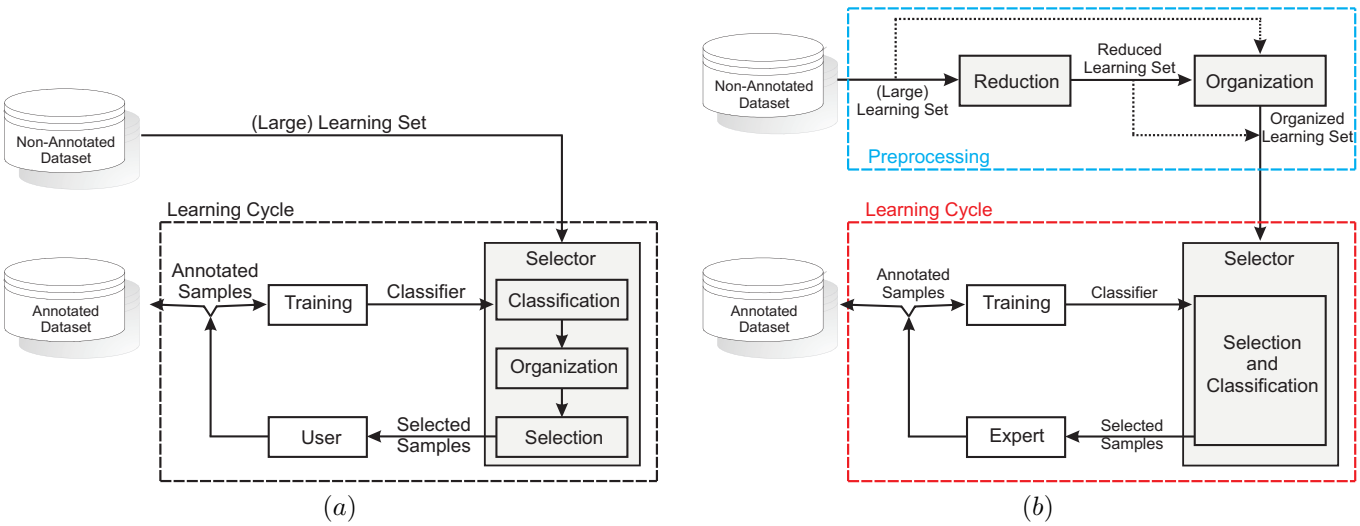


Fig. 1. Pipelines for active learning: (a) traditional paradigm and (b) proposed one.

protozoan cysts, helminth eggs, and larvae in Brazil. Their descriptors include shape, color and texture features weighted by optimization.

Therefore, the resulting datasets consist of images of those objects and their descriptors. The first dataset (d_1) contains 1,944 parasites (without impurities). The second dataset (d_2) has 5,948 samples, of which 1,944 are parasites and 4,004 are impurities. Both datasets were segmented by the system and carefully labeled by an experienced parasitologist expert. The third dataset (d_3) consists of over 140,000 unlabeled samples, unbalanced classes, absence of classes, and considerably higher number of impurities, which better reflects the circumstances present in an actual laboratory routine. The first two datasets were used to develop the method and the third one for final validation by an expert in Parasitology.

III. Thesis Contributions

The contributions of the Ph.D. thesis [4] consist of the proposal of a novel active learning paradigm and the development of new active learning methods associated with this paradigm. The proposed methods differ in the strategy used for data reduction and/or organization and the strategy used to select the best samples for expert supervision.

A. Data Reduction and Organization Paradigm

The main contribution of the thesis consists of a novel **D**ata **R**eduction and **O**rganization **P**aradigm (DROP) [8] for active learning, in order to select, more efficiently and effectively, a considerably lower number of the most informative samples to train a classifier under an expert's supervision. The major difference and advantage presented by the proposed paradigm (Figure 1b) is that the non-annotated dataset undergoes a priori reduction and/or organization, so that the selection does not require it to be entirely reprocessed at each learning iteration, unlike traditional active learning paradigms. As far as we know, the developed active learning strategies are unique in the sense that their data organization occurs only once (in a *unsupervised preprocessing phase*).

The proposed active learning strategies iteratively seek to select the most informative samples based on the synergy and current knowledge of both expert and classifier. The classifier actively participates in its learning process by classifying and supporting the selection of samples. During the *learning process*, one sample (at a time) on the ordered set is labeled by the current classifier, and the sample is selected if it receives the label that satisfies the given selection criterion. In general, the learning set is organized in pairs of samples such that, among the most difficult ones, the possibility of selecting sample pairs from distinct classes for annotation will be higher than pairs from the same class. Note that the classifier does not label all samples in the dataset. Both phases, classification and selection, are performed alternately until a desired number of samples per iteration is reached. Once selected, these samples are displayed to the expert for verification of the assigned labels and correction of the misclassified ones. Samples with expert-verified labels are then incorporated into the training set. As the classifier improves throughout the iterations, the expert's effort is increasingly reduced. The expert can direct the final classifier to annotate the remaining of the dataset when an acceptable accuracy has been reached, i.e., whenever the measured accuracy remains stable or reaches a sufficiently high level for the given application. Different classifiers could certainly be used, taking into account the time constraints of both classifier and application, as proposed in [9].

In the proposed paradigm, non-annotated samples can also be included in the training set to design a more effective classifier by *active semi-supervised learning* [10].

B. Active Learning Strategies

DROP aims to select unlabeled samples from all classes for expert annotation at the first learning iteration and then the most informative (hardest) samples for classification at

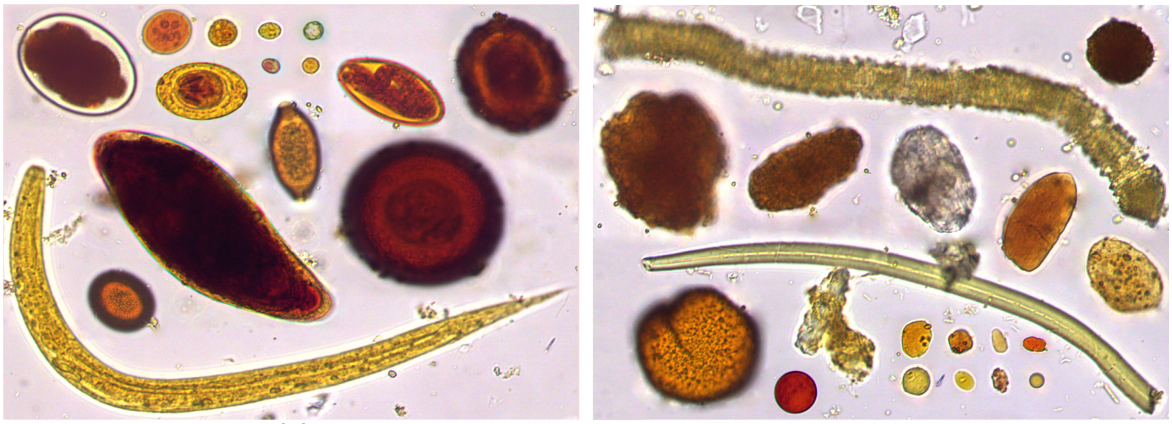


Fig. 2. Examples of image samples in the datasets. (a) from each class of parasites. (b) from impurities.

the subsequent iterations. Being a paradigm, it can be implemented with different strategies [11], [8], [10], [4], [12], [13], [14] for the reduction, organization and selection processes. Generally, the adopted strategies are related to graph-based clustering. After all, important information can be retrieved from clustering. Samples located at the center of clusters (root samples) are more likely to cover all classes and are good candidates to be selected first for manual annotation. Samples in the same cluster are likely to have the same label. Many of them should not be selected, avoiding redundant samples in order to accelerate active learning.

Cluster roots and boundary samples between distinct clusters, which form a reduced learning set, allow us to select the most informative samples earlier for training the classifier. After applying our boundary reduction strategy [14], as the Cluster-OPF-Rand instance, which performs a significant downsizing (by up to fifth percent, in our experiments) of the learning set, it is important to organize the remaining samples in a prioritized way, so that the most informative ones are more readily available for selection.

We proposed a **D**ecreasing **B**oundary **E**dges (DBE) organization strategy [13], in order to effectively arrange the samples of the reduced set. DBE organizes the reduced set based on the decreasing weight order of its boundary edges. The idea of prioritizing the largest edges formed by boundary samples is justified by those samples being, more likely than not, of different classes.

We also proposed a **M**inimum-**S**panning **T**ree **B**oundary **E**dges (MST-BE) organization strategy [8]. In order to increase the possibility of selecting boundary samples from distinct classes in the reduced set, the strategy interprets this set as a complete graph weighted by the distance between samples in the feature space, computes a Minimum Spanning Tree (MST) on it, and organizes the MST edges by decreasing weight order. The organization of the boundary set in decreasing order of distance between samples on its MST assumes that samples from the same class are usually the closest ones, and hence, they will be placed at the end of the resulting (organized) sample list, increasing the possibility of selecting samples from distinct classes sooner. Given that boundary edges with lower weights are more likely to be in the same class, MST-BE allows us to prioritize samples connected by edges with higher weights and classified in distinct classes during the selection strategy for expert annotation/verification.

The MST-BE strategy presents a better organization, selecting the more informative samples than the DBE strategy. Therefore, MST-BE was employed in the **A**ctive **S**emi-**S**upervised **L**earning (ASSL) [10], which is a novel integration of semi-supervised learning (proposed by [15]) and a priori-reduction and organization criteria (proposed by [8]) for active learning.

In the real problem of diagnosis of parasites, impurities are exceedingly abundant, form several clusters in the feature space, and are quite similar to some species of parasites (see Figure 2). Besides, there are unbalanced classes and absence of some classes, resulting in a major challenge for existing methods. In this context, under a scenario with the presence of the fecal impurity class, the proposed strategies with a reduction process was found to be considerably less effective. The data reduction can discard crucial samples for the learning process. In this case, it is important to exert care since some parasite species and/or impurities may be out of the cluster border.

Therefore, we also investigated a more robust solution for when there is the presence of a diverse class (such as impurities in the diagnosis of parasites), which early on organizes the data without discarding any of them. We proposed a new active learning strategy, called **R**oot **D**istance-**B**ased **S**ampling (RDS) [11] that pre-organizes the data and then properly balances the selection of diverse and uncertain samples for training. Data organization relies on clustering, followed by the sorting of the samples within each cluster based on their distance to their representative (root) sample. Selecting samples from the ordered list of each cluster, gives us a greater diversity. Selecting samples from each cluster according to the corresponding ordered list so long as their classification does not

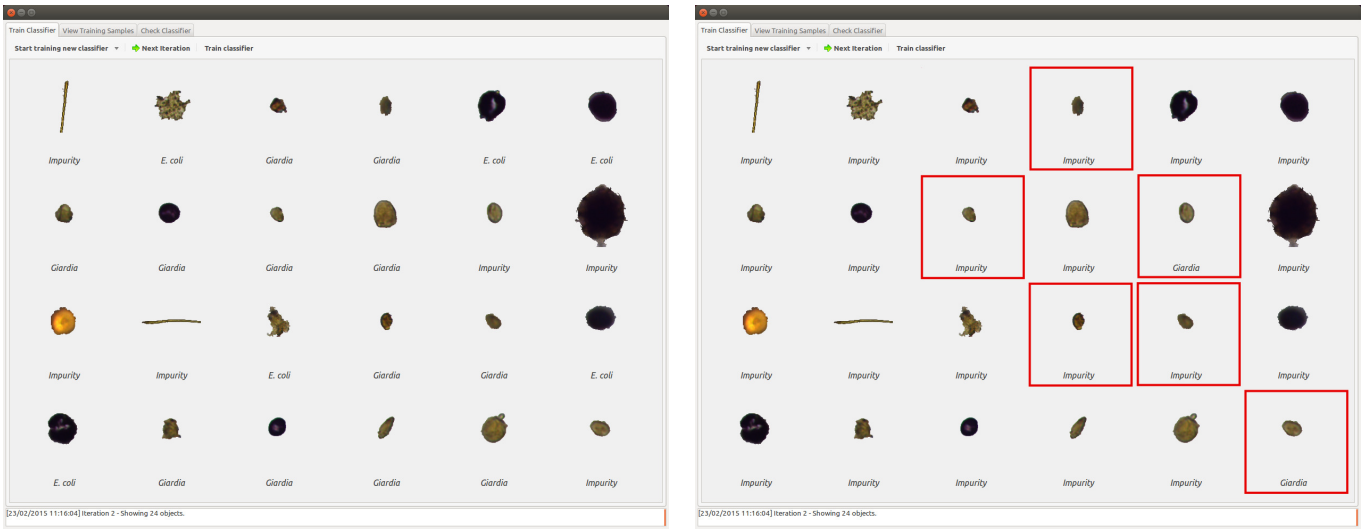


Fig. 3. Screen shots of the user interface, as used by the parasitologist to verify the label of selected objects: (a) images with labels given by the classifier. (b) images with labels corrected/confirmed by the parasitologist. Note that, *Giardia duodenalis* and some impurity components are difficult cases for class discrimination, as indicated by red squares.

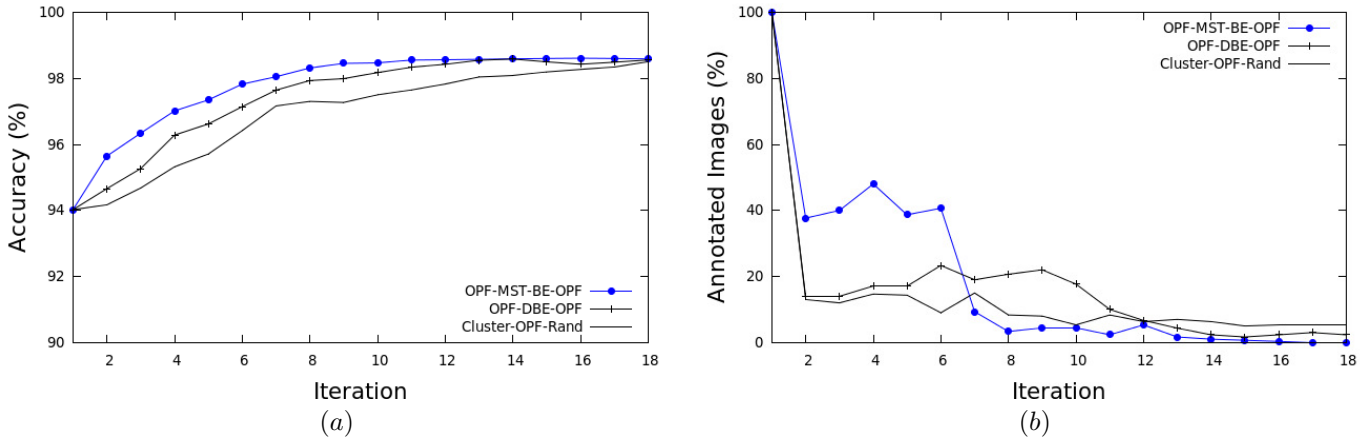


Fig. 4. Comparison between Cluster-OPF-Rand, DBE, and MST-BE methods using the OPF methodology (clustering and classification techniques) on the Parasites dataset (d_1). (a) Mean accuracy of the methods on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

match the class of the corresponding root, offered us the most difficult (most uncertain) samples for classification. When this condition is not satisfied, RDS selects uncertain samples by their decreasing distance to the cluster's root. Figure 3 shows the images with the labels given by the classifier (Figure 3a) and the same images with labels corrected/confirmed by the expert (Figure 3b), respectively. It is possible to observe some difficult cases for class discrimination.

IV. Experimental Results

First, we evaluated the methods using two versions of the labeled dataset, with and without impurities. The unlabeled dataset with 141,059 samples was used only to validate the best method by the parasitology specialist.

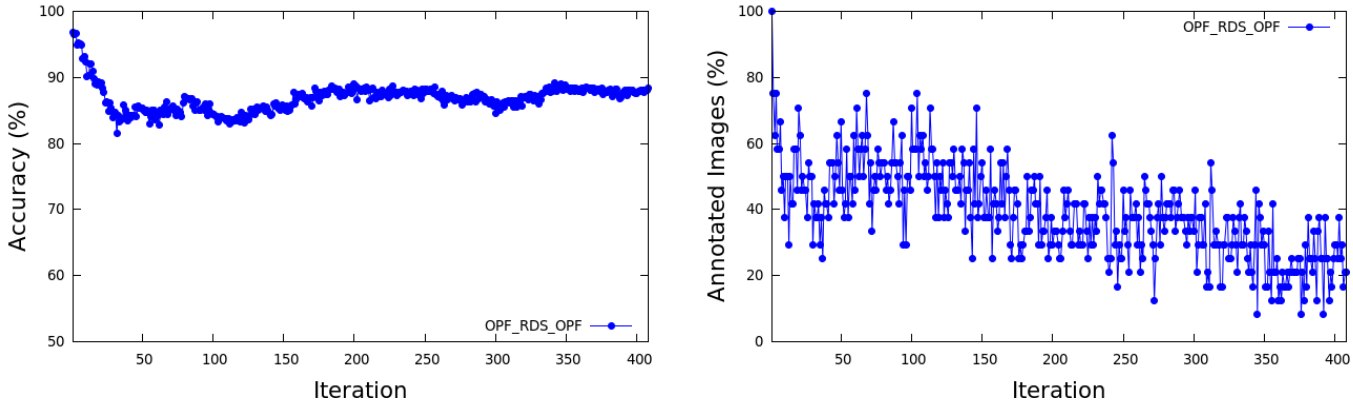
A. Main results of the developed methods

Initially, we present a comparison between the proposed methods (Cluster-OPF-Rand, DBE, and MST-BE). To compare the effectiveness of each method, we considered the accuracy measured (on an unseen test set obtained from the dataset) throughout the learning iterations as well as the percentage of annotated images in each iteration, using the first (d_1) dataset without impurities. Figure 4 shows that these methods were advances throughout this research towards an automatic classification of human intestinal parasites.

MST-BE with both techniques based on optimum-path forest presented the best results in a scenario without impurities. However, this is important to evaluate the robustness of the method with respect to the presence of such diverse class. Therefore, using the second dataset (d_2) with impurities, RDS's performance (accuracy on an unseen test set) was

TABLE I
MEAN ACCURACIES ± STANDARD DEVIATIONS OF THE METHODS ON THE PARASITES DATASET (d_2) WITH IMPURITIES.

Methods	OPF_ MST-BE_ OPF	OPF_ MST-BE_ SVM	Kmeans_ MST-BE_ OPF	Kmeans_ MST-BE_ SVM	OPF_ RDS_ OPF	OPF_ RDS_ SVM	Kmeans_ RDS_ OPF	Kmeans_ RDS_ SVM	Al- SVM	Rand_ OPF
accs	89.18%	85.96%	83.19%	81.40%	**91.58%**	90.27%	87.86%	84.90%	77.93%	74.07%
std dev	1.18±	1.72±	1.51±	1.83±	**0.90**±	1.79±	1.50±	1.53±	1.61±	2.10±



(a) (b)

Fig. 5. Results of the practical experiment performed by the parasitologist using OPF_RDS_OPF on the Parasites dataset (d_3) with impurities: (a) Mean accuracy of the method on unseen test sets. (b) Annotated images, as a percentage of the displayed samples, in each iteration.

validated against two baseline active learning methods: Al-SVM [16], which selects samples from the entire learning set at each iteration using an SVM classifier, and the most competitive one, MST-BE [8]. We also compared RDS with a random method in which samples were randomly selected from the entire dataset. For clustering (used for data organization), we evaluated the OPF and $k$means techniques. For classification (used for training sample selection), we used the SVM and OPF classifiers.

In order to facilitate the comparison among methods, when applicable, they were labeled as a triple, consisting of the clustering method, active learning method, and classification method, separated by an underscore character. The methods were denoted as Kmeans_RDS_OPF, OPF_RDS_OPF, Kmeans_RDS_SVM, OPF_RDS_SVM, Kmeans_MST-BE_OPF, OPF_MST-BE_OPF, Kmeans_MST-BE_SVM, OPF_MST-BE_SVM, and Al-SVM.

Table I shows the results of the comparisons among the methods for the case of the dataset (d_2) with impurities. The active learning methods (RDS and MST-BE) are superior to Al-SVM and Rand_OPF methods, for both OPF and $k$means clustering techniques, as well as for both OPF and SVM classifiers. However, RDS outperformed MST-BE. In general, the RDS method (using both OPF clustering [17] and classifier [18]) had the best performance (achieving higher accuracies and decreasing the number of annotated images earlier, as well as presenting shorter learning times) in the presence of impurities. Therefore, we selected OPF_RDS_OPF method for evaluation by the specialist on the chosen realistic dataset d_3 (see Section IV-B).

B. Results of the best approach for the diagnosis of parasites

In this section, we present the results of the OPF_RDS_OPF method. A 10-fold cross-validation was calculated in the training set to predict the accuracy per iteration and guide the expert when to stop the learning process. To evaluate the final accuracy, a random subset of the remaining unlabeled objects was selected and automatically annotated by the final classifier. These objects were evaluated by the expert, who indicated the classification errors to compute the final acuracy.

Figure 5 shows the 10-fold cross-validation average accuracy and the percentage of annotated images in each iteration. We can see that the predicted accuracy started high and decreased when new species were detected by the expert, until it stabilized within a small range. After 408 iterations (24 images per iteration), the expert verified 9,792 objects (6.9% of the dataset) and corrected the label of 3,796 objects (38.7% of the selected objects). The expert decided to stop the learning process when the mean accuracy by cross-validation on the labeled samples stabilized between 87% and 88%.

In order to evaluate the real accuracy of the final classifier, we created a random subset with 6% of the remaining unlabeled samples (7,870 objects). The classifier achieved 87.2% of accuracy on the random subset, matching the accuracy predicted during the training phase. This is a remarkable result, considering the low sensitivity rates from the traditional diagnosis procedure, based on visual analysis (48.3% up to 75.9%) [5], as well as taking into account that the expert verified only 6.9% of 141,059 objects.

Table II shows the mean accuracies per class. The classification performance for *Entamoeba histolytica/E. dispar cysts*

TABLE II
MEAN ACCURACIES FOR EACH CLASS, USING OPF_RDS_OPF IN A REALISTIC SCENARIO (DATASET d_3).

Species	Accuracies
Entamoeba histolytica / E. dispar	60.16%
Giardia duodenalis	72.83%
Entamoeba coli	86.75%
Endolimax nana	84.82%
Iodameba bütschlii	47.50%
Blastocystis hominis	79.03%
Ascaris lumbricoides	94.40%
Enterobius vermicularis	91.43%
Ancylostomatidae	92.24%
Strongyloides stercoralis	91.96%
Trichuris trichiura	95.15%
Hymenolepis nana	93.95%
Hymenolepis diminuta	95.97%
Taenia spp.	96.48%
Schistosoma mansoni	91.38%
Impurities	80.36%

and *Iodameba bütschlii cysts* is likely to improve whenever we are able to increase the number of samples from these classes.

V. CONCLUSION

In this research, we proposed original solutions to deal with a real application, the automated diagnosis of intestinal parasites. The methods were published in international conferences [10], [12], [13], [14] and top-tier international journals [11], [9], [8].

The major contribution of the Ph.D. thesis [4] was a novel active learning paradigm that affords interactive response time and verification of a considerably smaller part of the dataset, allowing its application to large datasets. We also proposed active learning strategies that were extensively evaluated with different types of unsupervised and supervised classifiers as well as with baseline learning strategies and using datasets from distinct application domains, of different sizes, and with feature spaces of various dimensions and classes, such as: image segmentation, forest cover type, handwritten digits, faces, cowhide, and image annotation from the real application of diagnosis of parasites.

The experiments performed on these datasets show that the proposed strategies outperform the baseline ones, requiring only a few iterations to identify samples from all classes and achieve high accuracy with less expert involvement. Moreover, the most suitable strategy for a real application — the automated diagnosis of human intestinal parasites — was evaluated and validated by an experienced expert in parasitology using a realistic scenario. We have demonstrated the good performance of our strategy, reaching average accuracies (about 90%) higher than those (between 48.3% and 75.9%) currently practiced in public and private clinical laboratories, which use conventional parasitological techniques and visual analysis of slides. This is a remarkable result, specially when we take into account that the expert verified only 6.9% of 141,059 samples. We believe that our solution is a very relevant contribution to the area of clinical parasitology.

ACKNOWLEDGMENT

We would like to thank FAPESP, CNPq, CAPES, Fundação Araucária and UTFPR for their financial support.

REFERENCES

[1] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.

[2] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos, "Scalable active learning for multi-class image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2259–2273, 2012.

[3] Y. Fu and X. Zhu, "Optimal subset selection for active learning," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. AAAI Press, 2011, pp. 1776–1777.

[4] P. T. M. Saito, "Active learning with applications to the diagnosis of parasites. PhD Thesis, Institute of Computing, University of Campinas (IC-UNICAMP). Available: http://www.bibliotecadigital.unicamp.br/document/?code=000931972," p. 72, 2014.

[5] J. F. Gomes, S. Hoshino-Shimizu, L. C. S. Dias, A. J. S. A. Araujo, V. L. P. Castilho, and F. A. M. A. Neves, "Evaluation of a novel kit (tf-test) for the diagnosis of intestinal parasitic infections," *Journal of Clinical Laboratory Analysis*, vol. 18, no. 2, pp. 132–138, 2004.

[6] C. T. N. Suzuki, J. F. Gomes, A. X. Falcão, J. P. Papa, and S. Hoshino-Shimizu, "Automatic segmentation and classification of human intestinal parasites from microscopy images," *IEEE Transactions on Biomedical Engineering (TBME)*, vol. 60, no. 3, pp. 803–812, 2013.

[7] A. X. Falcão, J. Stolfi, and R. d. A. Lotufo, "The Image Foresting Transformation: Theory, algorithms, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[8] P. T. M. Saito, P. J. de Rezende, A. X. Falcão, C. T. N. Suzuki, and J. F. Gomes, "An active learning paradigm based on a priori data reduction and organization," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6086–6097, 2014.

[9] P. T. M. Saito, R. Y. M. Nakamura, W. P. Amorim, J. P. Papa, P. J. de Rezende, and A. X. Falcão, "Choosing the most effective pattern classification model under learning-time constraint," in *PlosOne*, 2015, pp. 1–23.

[10] P. T. M. Saito, W. P. Amorim, A. X. Falcão, P. J. de Rezende, C. T. N. Suzuki, J. F. Gomes, and M. H. de Carvalho, "Active semi-supervised learning using optimum-path forest," in *22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 3798–3803.

[11] P. T. M. Saito, C. T. N. Suzuki, J. F. Gomes, A. X. Falcão, and P. J. de Rezende, "Robust active learning for the diagnosis of parasites," *Pattern Recognition*, pp. 1–12, 2015.

[12] J. E. Vargas, P. T. M. Saito, A. X. Falcão, P. J. de Rezende, and J. A. dos Santos, "Superpixels-based interactive classification of very high resolution images," in *XXVII SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2014, pp. 173–179.

[13] P. T. M. Saito, P. J. de Rezende, A. X. Falcão, C. T. N. Suzuki, and J. F. Gomes, "A data reduction and organization approach for efficient image annotation," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*, 2013, pp. 53–57.

[14] ——, "Improving active learning with sharp data reduction," in *WSCG Communication Proceedings of 20th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2012, pp. 27–34.

[15] W. P. Amorim, A. X. Falcão, and M. H. de Carvalho, "Semi-supervised pattern classification using optimum-path forest," in *XXVII SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2014, pp. 111–118.

[16] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[17] L. M. Rocha, F. A. M. Cappabianco, and A. X. Falcão, "Data clustering as an optimum-path forest problem with applications in image analysis," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 50–68, 2009.

[18] J. P. Papa, A. X. Falcão, V. H. C. de Albuquerque, and J. M. R. S. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognition*, vol. 45, no. 1, pp. 512–520, 2012.