Active Labeling: Application to Wireless Endoscopy Analysis

Petia Radeva[1,2] Michal Drozdzal[1,2] Santi Segui[1,2] Laura Igual[1,2]
Carolina Malagelada[3] Fernando Azpiroz[3] Jordi Vitria[1,2]
[1] Department of Applied Mathematics and Analysis, Universitat de Barcelona, Spain
[2] Computer Vision Center, Barcelona, Spain
[3] 3 Digestive System Research Unit, Hospital de Vall d'Hebron, Spain

INVITED PAPER

Abstract

Today, robust learners trained in a real supervised machine learning application should count with a rich collection of positive and negative examples. Although in many applications, it is not difficult to obtain huge amount of data, labeling those data can be a very expensive process, especially when dealing with data of high variability and complexity. A good example of such cases are data from medical imaging applications where annotating anomalies like tumors, polyps, atherosclerotic plaque or informative frames in wireless endoscopy need highly trained experts. Building a representative set of training data from medical videos (e.g. Wireless Capsule Endoscopy) means that thousands of frames to be labeled by an expert. It is quite normal that data in new videos come different and thus are not represented by the training set. In this paper, we review the main approaches on active learning and illustrate how active learning can help to reduce expert effort in constructing the training sets. We show that applying active learning criteria, the number of human interventions can be significantly reduced. The proposed system allows the annotation of informative/non-informative frames of Wireless Capsule Endoscopy video containing more than 30000 frames each one with less than 100 expert "clicks".

Keywords: interactive labeling, active learning, WCE

I. Introduction

Today, a huge amount of real applications apply machine learning techniques (e.g. pedestrian recognition, speech recognition, biometrics identification, information extraction, document classification, image retrieval, etc.). In order to obtain a highly precise classifier performance the supervised machine learning should be based on thousands of training samples. The best strategy to build a realistic model of a given class of samples is to collect as much training samples as possible. In most applications, achieving high amount of data is not a problem. The real problem comes when data sets should be labeled as positive or negative examples. In some cases, labeling assumes relatively low cost like the spamm annotations or ranking the movies by people preferences. However, in many other applications the labeling is not a trivial task. Let us imagine all kind of imaging of coronary vessels (computer tomography, magnetic resonance angiography, intravascular ultrasound images, optical coherence tomography, etc.). Obtaining high amount of images usually is not a problem since any clinical imaging easily acquires hundreds or thousands of images. The real problem comes when these images should be labeled according to the presence of a given lesion e.g. atherosclerotic plaque deposit. Such labeling makes the learning process tedious, subjective and time consuming and highly expensive taking into account that lesions or other clinical objects should be detected, identified and delineated by highly trained experts of the application field. In this context, time is expended in two different processes: 1) the labeling process, which generally needs human intervention in processing thousands of images, and 2) the training process, which in some cases exponentially should increment computational resources as more data are obtained.

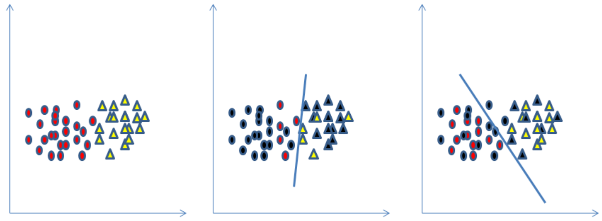

Figure 1. An illustrative example when active learning can help obtaining optimal classifier with small number of examples: hidden labels (left), guided labeling by active learning strategy (center) and random sampling of training set labeled by the user (right).

Given a sequential data source, the problem has been alleviated by integrating online learning methodologies and a controlled choice of training samples to be added into the training process. Given a training set, an "intelligent" system should add to it only those data samples that were not represented, or under-represented, in the previous version of the set. Therefore, the training set should be enlarged by those new data samples that enrich the representability of the classification sets distributions while avoiding unnecessary samples redundance. Even in this situation, the labeling process is still a problem. To overcome this problem, we

978-1-4673-2362-8/12/$31.00 ©2012 IEEE

Figure 2. The wireless video capsule

should use criteria to optimize the labeling process, i.e. to minimize the amount of human interaction during the training set construction. The main challenge is to decide which samples should be considered by the experts in order to optimize the training process. Active learning systems have the purpose to overcome the labeling bottleneck asking queries and deciding which unlabeled samples should be considered by the experts. Active learning is of high interest in many machine learning applications where data used to be abundant but labels are expensive or very scarce. Thus, we assume to have a big collection of data points, each of them belonging to a given class although not known from the beginning. Taking into account that usually labeling each sample has some cost, we are interested in finding a classifier that will accurately map points to label without spending too much. The original idea of active learning is that if the process of creating the training set is guided according to some criteria, the task of the learner can be made easier. Fig. 1 shows a toy example of the effect of Active learning. Let us imagine we have two classes represented in the left figure shown with their labels although at the beginning most of the labels are unknown. If we choose the labels given in the right image and train a classifier, it is easy to construct a suboptimal one. However, if we choose samples that help the classifier to decide the right boundary in the frontier between the clusters (the central image), we will get better classification performance with relatively small number of examples. A complete and intuitive overview of active learning techniques can be found in [13].

A possible strategy to optimize the labeling process is to embed a model of the data into a continuous labeling process, where the system predicts a label of samples previously estimated most critical for the system and the human operator should confirm the label for them. The human operator faces two possible decisions: to accept the model proposal or to change the label of the sample. In practice, these two choices have a non symmetric cost for the human operator: accepting the model proposal can be efficiently implemented with a low cognitive load for the operator, while changing a label has a larger cost. This effort can be measured by the number of interventions the operator should perform during the process. In this paper, we illustrate the advantage of applying active learning techniques to construct the training set for informative frames classification in wireless endoscopic images.

Wireless video capsule endoscopy was presented in 2000 [1] by Given Imaging Ltd. [2] and it is still now the latest evolution in gastrointestinal endoscopy and the only one that allows the whole visualization of the small bowel. Due to its numerous clinical advantages, it has rapidly become a wide-spread clinical routine and its use has been proposed for the categorization of diverse pathologies, such as Crohn's disease, tract bleeding and polyp search [3], [4], only to cite a few. Wireless Capsule Endoscopy (WCE) image analysis (see Fig. 2) is a clear scenario where these problems arise. The WCE consists of a small ingestible pill (11mm x 26 mm, 3.7g.) which contains a camera and a full electronic set which allows the radio frequency emission of a video movie in real time. This video, showing the whole trip of the pill along the intestinal tract, is stored into an external device which is carried by the patient. These videos can easily have duration up to 8h, what leads to capsule capturing of a total of 7.200 to 60.000 images. WCE videos have been widely used to create computer-aided systems to differentiate diverse parts of the intestinal tract like esophagus, stomach, duodenum, jejunum-ileum and cecum [5], to measure several intestinal disfunctions [6] and to detect different organic lesions (such as polyps [7], bleeding [8] or general pathologies [9]).

From point of view of computer-aided systems, researchers have focused their efforts on trying to tackle the inherent drawbacks associated to the video screening stage of capsule endoscopy videos: long time needed for visualization -in order to make a decision, a specialist should look at more than 40.000 images, i.e., the specialist should spend several hours on a specific clinical case-, potential subjectivity of the observer due to fatigue, presence of intestinal contents which hinders the proper visualization of the intestinal walls, etc. In the recent literature, we can distinguish three general lines of research with respect to the aim of each proposed system, namely: 1) The reduction of the final time needed for visualization, or adaptive control systems for video display [10]; 2) The characterization of different lesions in the gut such as polyps [7], bleeding [8] or general pathologies [9]; and 3) the differentiation of the diverse parts of the intestinal tract like esophagus, stomach, duodenum, jejunum-ileum and cecum [5]. In addition to the former, in the last years different authors have payed attention to the study and characterization of specific events of intestinal motility, such as intestinal contractions or motor activity [6].

One of the important problems to diagnose intestinal motility disorders is based on determining the quantity and distribution of informative frames that can vary from a few to several thousands of images per video. A common stage in all research lines is the discrimination of informative frames from non-informative frames. Non-informative frames are defined as frames where the field of view is occluded. Mainly, the occlusion is caused by the presence of intestinal content, such as food in digestion, intestinal juices or bubbles (see Fig. 3). The ability of finding non-informative frames is important since: 1) generally, it helps to reduce time of video analysis, and 2) since the majority of non-informative frames are frames

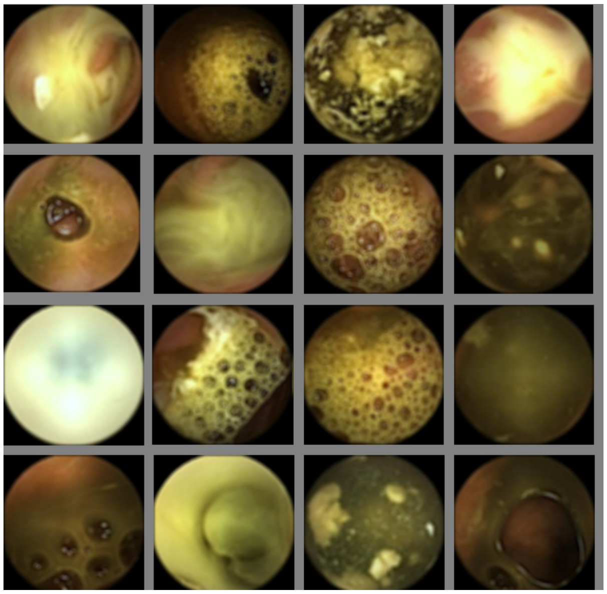

Figure 3. Non informative endoscopy frames.

with intestinal content which information can be used as an indicator for intestinal disfunctions [11].

The main strategy in the non-informative frames search is the application of machine learning techniques in order to build a two-class classifier. Generally, non-informative frames are characterized by the color information [12]. Robust classifiers can be built when the training set is representative of the data population. The problem that occurs when dealing with WCE is the high color variability of non-informative frames in different videos. It is very probable that, as new videos become available, a new video with significantly different intestinal content color distribution will be added to the training set. A naive approach for the labeling process could mean the manual annotation of up to 50.000 frames per video.

Active learning provides powerful techniques for interactive labeling that allows to reduce significantly the number of human interventions during the labeling process of a new video. The process is applied to the labeling of non-informative frames for WCE image analysis. In the active learning algorithm the key idea is that machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which is learns [13], [14]. However, the goal of the system is basically to enlarge the training set while minimizing human intervention, instead of correcting the errors of the classification process like in classical active learning approaches [13]. This is done by finding an optimal classifier adaptation scheme for non-represented frames in the original training set.

The rest of the paper is organized as follows: Section 2 overview the Active learning techniques, Section 3 introduces the interactive labeling method, Section 4 describes the application of active learning to for WCE labeling. In Section 5, we present experimental results, and finally, in Section 6 we expose some conclusions and remarks on future research.

II. Active Learning

In general, the active learning problem deals with how to define a criterion on the unlabeled data to be queried so that the training process is optimized with respect to some aspect (learner performance (Fig. 4 (left)), training speed (Fig. 4 (right)), labeling cost, etc.). There are several questions that arise in Active learning: what is the optimal set of examples to give to the classifier?! What is the minimal set of examples to be labeled in order to achieve optimal classification results. Usually, there are two directions to optimize the learning process: a) by choosing the points to query that shrink the space of possible classifiers as much as possible, or b) by exploiting the structures of data distributions to find clusters of unlabeled data. In some cases, the techniques may depend on the scenarios where the data live.

A. Scenarios for input spaces of query samples

There are three different scenarios in which the learning system will be able to ask queries to the user. The main settings are: a) membership query synthesis, b) stream-based selective sampling, and c) pool-based sampling. In the membership query synthesis, the learner request labels for instances in the input space including queries the learner generates de novo, instead of those generated or sampled from some underlying natural distribution. In this class of problems, the query can be synthesized and thus it is applied to regression problems like the case of learning to predict the absolute coordinates of a robot hand given the joint angles of its mechanical arm as inputs [18].

In the stream-based selective sampling, the main idea is that to obtain an unlabeled instance is free and/or inexpensive so it is sampled from the actual distribution and then the learner decides whether to query its label. When the distribution is uniform, selective sampling behaves as membership query learning. However, if the distribution is non-uniform and unknown, the queries will be sensible since they will come from a real data distribution. Real applications of stream-based selective sampling include: learning ranking functions for information retrieval [21], word sense disambiguation [20] or part-of-speech tagging [19]. There are several criteria applied to decide whether to query a sample or not. Some authors estimated an ”informativeness measure” or ”query strategy” and do a biased random decision so that more informative instances are likely to be queried [19]. Another strategy is to explicitly compute the region of uncertainty by for example setting a minimum threshold on the informativeness measure which defines the region. Only instances above the threshold are requested. As an alternative, criterion based on disagreement of models of the same model class but with different parameter settings disagree on the unlabeled instances, the instance is considered to belong to a uncertainty region [22].

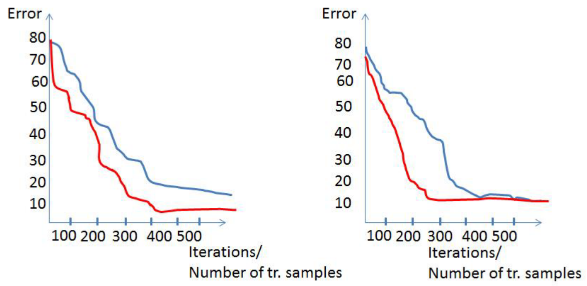

Figure 4. Effect of Active learning: a) improvement of learner performance (left), b) Improvement of training time instead of final performance (right). In both figures, the blue line shows the error reducing when the learner is fed by training samples selected by random sampling; the red line shows the error reducing when the learner is fed by training samples chosen by active learning strategy.

In the case of pool-based sampling, we assume to have a large closed pool of available samples where only a small part is labeled. Queries are done on the unlabeled samples usually in a greedy procedure according to an informativeness measure that evaluates all samples from the pool. Several applications are known from the bibliography like: image classification and retrieval [23], video classification and retrieval [25] or speech recognition [24].

B. Criteria for the choice of the query samples

Most active learning scenarios are based on evaluating in some way the informativeness measure of unlabeled instances that can be generated or sampled from a given distribution. One of the simplest and most popular criterion for sample queering is the uncertainty sampling. In this strategy, the active learner choose the sample to query that is the most uncertain if it should be labeled by the learner. This approach is easily applicable for probabilistic learning models. For example, in case of binary problems, the samples will be queried if their posterior probability is close to 0.5. In case of multi-class problems, the least confident samples will be queried first [26]. In order to take into account the distributions of the most probable label as well as the remaining distributions, in [27] labels are chosen on the criterion on margin sampling i.e. that have minimal difference in the first two probable labels. A more general strategy for uncertainty sampling is based on entropy that moreover, easily generalizes to probabilistic multi-label classifiers and probabilistic models for structured instances like sequences [26].

The query-by-committee is a more theoretically-motivated query selection framework that assumes a committee of learners that have been trained on the current labeled set but represent competing hypotheses. After voting by all learners, the sample with the highest disagreement of the committee is chosen to be queried [28]. The fundamental message beyond the formal definition of this criterion is to minimize the vector space that is the set of hypotheses consistent with the current labeled training set. In order to implement a Query-by-Committee selection algorithm, it should be possible to construct a committee of models representing different regions of the version space as well as to define a measure of disagreements among different members of the commitee. For example, in [30] the Jensen-Shannon divergence has been used to measure the disagreement. In [29], a pool-based margin strategy for Support Vector Machines is used to minimize the version space directly.

A very different approach is presented by the expected model change criterion. In this case, a decision-theoretic approach is applied to decide the sample that would cause the biggest change of the model if its label were known. A very interesting example of this approach is given by the expected gradient length criterion for discriminative probabilistic model classes. Given that the discriminative probabilistic models are trained by gradient-based optimization, the change expected for the model can be measured by the training gradient length that is the vector used to reestimate the parameter values. The learner queries those instances which when labeled and added to the training set, would lead to the biggest magnitude of the gradient. This approach gives very nice final results as soon as the feature space and the set of labeling are not large. Moreover, if one of the features has too large values it would result in a high gradient magnitude that would provoke informativeness overestimation. Parameter regularization is recommended to avoid such negative effects like those applied in [31].

The expected error reduction criterion is another decision-theoretic approach that focuses on the reduction of the generalization error rather than the model change. Estimating the expected future error of the enriched model with the sample candidate on the remaining unlabeled instances, the learner queries the sample with the minimal expected future error or risk. The formal application can be expressed in terms of minimizing the 0/1- loss or the expected log-loss that is equivalent to reducing the expected entropy over the unlabeled set of samples. It can be shown that in some way this criterion is equivalent to maximization of the mutual information of the output variables or the expected information gain of the query. This criterion was very successfully applied in a combination with a semi-supervised learning approach leading to a very significant improvement over random and uncertainty sampling [33]. The authors in [32] employ a variant that biases the expectation toward the most likely label using uncertainty sampling as a fallback strategy when the learner provides an unexpected labeling. However, this criterion is highly expensive from computational point of view. The expected error should be estimated over all unlabeled samples as well as the new model should be retrained for each possible query labeling. As a result, this framework is applicable to mainly binary problems and simple models like Gaussian random fields [34]. Resorting to Monte Carlo sampling from the pool [33] to reduce the unlabeled set sampling as well as applying approximate training techniques [32] can significantly make more practical the computational procedure.

Minimizing the expectation of a loss function is not a trivial task, in general it is not possible to perform it in a closed form. If we were able to reduce the generalization error indirectly by minimizing the variance of the output, we would be able to achieve a closed-form solution. For example, if we minimize the standard error of a regression problem, we can take advantage of the decomposition of the expected future error in terms of the expectation of the labeled set, expectation error over the conditional density and the expectation over both [35]. In other words, the expectation error can be reformulated as a three terms composition where the first term is basically noise independent on the model and training data, the second term is the bias representing the error, due to the model class itself and thus being invariant given a fixed model class, and the third term is the model variance which is expressing the remaining component of the learner squared loss with respect to the target function. Thus, by minimizing the variance guarantees the reduction of the generalization error of the model. To minimize the variance over its parameter estimates, an active learner can select the data that maximizes the Fisher information [36]. When there is only one parameter, the computation is an easy task. However, when there are K parameters of the model, the Fisher information takes form of K by K covariance matrix making difficult the optimization process. There are three main approaches in such cases: minimization of the trace of the inverse information matrix, of the determinant of the inverse matrix or of the maximum eigenvalue of the inverse matrix. The practical disadvantages of these variance-reduction methods refer to the high computational complexity. When the model has substantial number of parameters, estimating output variance requires inverting high-dimensional matrix for each new instance. This fact easily leads to intractable applications in real problems like natural language processing tasks. The authors in [37] apply principal component analysis to reduce the dimensionality of the parameter space. In [26] the authors approximate the matrix with its diagonal vector to invert it in linear time. However, this strategy is still significantly slower compared to other strategies like the uncertainty sampling.

If we aim to be less sensible to outliers, it is a good idea to focus on the entire input space rather than individual instances. For example, uncertainty sampling criterion could prefer to query labels that are just close to the learner boundary independently if they in some way are related to the rest of instances or not. If we take into account the unlabeled pool when estimating errors and variances, the estimated error reduction strategy would avoid these shortcomings. To this purpose, the authors in [26] propose a general density-weighted criterion. The main idea beyond is that the chosen instances to query should be at the same time informative as well as representative of the underlying distribution e.g. live in a dense region of the input space. In this context, we can find several works that successfully applied such criteria like those in [40] that construct sets of queries by clustering for batch-mode active learning with SVMs or a method that precomputes the density in order to select the next query achieving time to select the next query comparable to the base informative measure in uncertainty sampling.

An excellent alternative is presented in [38], [39] where the authors focus on exploiting natural clusters in the unlabeled data sets. The data are organized in hierarchical clusters from where samples are chosen to be queried according to the purity of the cluster. As a result, a tree prunning is constructed in terms of clusters of homogenous labels that are used as an optimal training set obtained by minimal human intervention for the classification problem. The important advantage of the method is that the algorithm does not make any assumption about the data distribution of data and labels, maintains valid estimates of the error induced by its current prunning and refines prunnings to drive down the error. Moreover, it allows to explore different clustering procedures and sampling strategies. Still, we should note that, in general, there is no clear characterization of when the active learning methods behave better than the straight random sampling and by how much [38]. All these methods open new directions for theoretical studies and practical applications that hide a lot of potential for optimizing the complex process of training set construction and optimal classification with a minimal and optimized user intervention. Hence, we expect to see in the near future a lot of theoretical advances and real application of active learning in fields that are strong users of machine learning techniques.

III. Interactive Labeling

Given the problem of labeling all informative frames in videos of thousands of frames each one, our purpose is to design an interactive labeling system that should allow, in an efficient way, 1) to detect frames that are not represented in the training set, 2) to obtain statistics of informative and non-informative frames that are not represented in the training set, and 3) to be able to iteratively increase the representability of the training set in an ”intelligent” way by reducing significantly the number of clicks related to manual labeling. To this aim, we propose the following algorithm for interactive labeling of a set of new images optimizing the user feedback:

1) Let be L a set of labeled data, M_1 a discriminative model trained on this set, U a set of all unlabeled data samples from a new video and C a criterion to select the most informative frames from U;
2) Select the subset of samples $N = \{x_j^N\}$ from U such that they are considered as under-represented by the labeled set L.
3) Evaluate the subset N with M_1, assigning a label l_j to every data sample x_j from N.
4) i=1;
5) **While** there are elements in N
Evaluate the elements of N with respect to the criterion C and get the set of n most informative samples $I \subset N$ (with the purpose of minimizing the expected number of expert clicks).
6) Delete the elements of I from N.

```
7) Present the samples from $I$ to the user with their associated label.
8) Get the user feedback (nothing for samples with correct labels, one click for each wrongly classified sample).
9) Update $L$ by adding the elements of $I$ and its user-corrected label.
10) Perform an online training step for $M_i$ by adding the elements of $I$ to the model, getting $M_{i+1}$.
11) Evaluate the elements of $N$ with $M_{i+1}$, assigning a label $l_j$ to every data sample $x_j$ from $N$.
12) $i = i + 1$;
    Endwhile
```

The critical step in order to minimize the number of clicks is to choose a good criterion for Step 6, since it represents the main strategy of choosing the order of presentation of the samples to be labeled by the user. To this end, we studied three different sorting policies for the elements of N. These policies are based on the following criteria: 1) to choose those elements that are far from the training data L and far from the boundary defined by M_i, 2) to choose those elements that belong to the most dense regions of N, and 3) to choose the elements in a random way.

More specifically, we define them in the following way:
Criterion 1 (C1)*: Distance of data to the classifier boundary and training data.* In this criterion, two measurements are combined: 1) The data are sorted from the farthest to the nearest distance with respect to the classifier boundary. This scheme assumes that the classifier, while proposing labels, will commit errors with higher probability for the samples that are far from the boundary than for the data that are relatively close to boundary. 2) The data are sorted from the farthest to the nearest with respect to the training data. This scheme assumes that the classifier, while proposing labels, will commit errors with higher probability for the samples that are far from the training set than for the data that are relatively close to the known data. A final sorting is performed in the data by adding the ranking indices of the two previously described schemes.
Criterion 2 (C2)*: Data density.* Each sample is sorted decreasingly with respect to a data density measure in its environments. This scheme assumes that the classifier should learn more quickly if we first provide samples from the zones with higher density. Data density can easily be computed as the mean distance to the k-nearest neighbors of the sample.
Criterion 3 (C3)*: Random order.* The order of presentation of the samples randomly determined.

IV. The Interactive Labeling System

The goal of the interactive labeling system is two-fold: 1) to detect, for each new video, the set of frames that are not represented in the training set, and 2) to label those frames with minimal user effort. To this end, we propose a system design with two main components(see Fig. 5):
1. A data density estimation method that allows fast local estimation of the density and distance of a data sample to other examples, e.g. from the training set (see Step 3 of the algorithm for interactive labeling).
2. An online discriminative classifier which allows to sequentially update the classification model M_i of thousands of samples (see Step 2, 10 and 11 of the algorithm for interactive labeling).

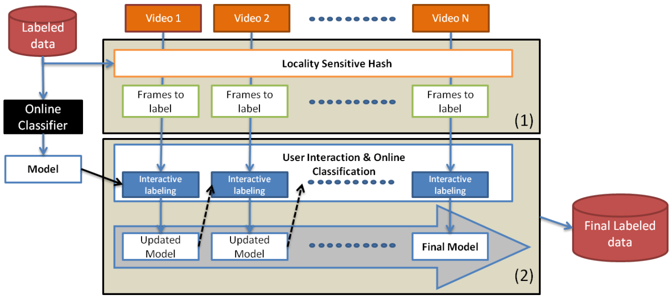

Figure 5. The interactive labeling system architecture with its two main components: 1) Detection of frames not represented in the training set and 2) Labeling of frames and model enlarging using online classifier method.

A. *Fast Density Estimation*

As previously commented, the local density of a data sample x_i with respect to a data set C can be easily estimated by computing the mean distance from x_i to its k-nearest neighbors in C. The simplest solution to this problem is to compute the distance from the sample x_i to every sample in C, keeping track of the ”k-best so far”. Note that this algorithm has a running time of $O(nd)$ where n is the cardinality of C and d is the dimensionality of samples.

Because of excessive computational complexity of this method for large data sets, we need a flexible method that allows from one side, effective measurements of characteristics of large data and, from the other side, introducing new unseen data into the training set for enlarging the data representation. An example of such flexible method is Locality Sensitive Hashing[15]. LSH allows to quickly find a similar sample in a large data set. The basic idea of the method is to insert similar samples into a bucket of a hash table. As each hash table is created using random projections over the space, several tables can be used to ensure an optimal result [15]. Another advantage of LSH is the ability to measure the density of data in given space vicinity by analyzing the number of samples inside the buckets (Fig. 6). In order to evaluate if the new sample improves the representation of the data set, the space density of the training set is estimated. If the new sample is in a dense part of the space then the sample is considered redundant and thus it is not considered in order to improve the training set. Otherwise, the sample is used to enlarge the training set.

The density D of the sample x is estimated according to the formula:

$$D(x, T_r) = \sum_{i=1}^{M} ||B_i|| \tag{1}$$

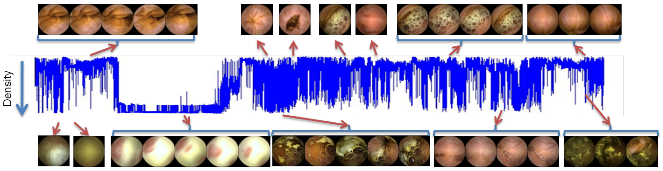

Figure 6. Example of training set density estimation for a test video using LSH. The images show the zones of high and low density with respect to given labeled set L.

where M is the number of hash tables, $||B_i||$ is the number of elements in the bucket where the new element x is assigned and T_r represents the training set. The subset of not-represented samples in the training set $N = \{x_1^*, ..., x_m^*\}$ from new unlabeled data U is defined as:

$$N = \{\forall x \in U : D(x, T_r) < T\} \quad (2)$$

where T represents a fixed threshold.

Note that (2) expresses the condition that the new samples fall in buckets with low density of the training set. That is if the new sample is in a dense part of the space then the sample is not considered to improve the training set. Otherwise, the sample is used to enlarge the training set.

B. Online Classifier

Taking into account that our classifier must be retrained with thousands of images/feature vectors of up to 256 components, using an online classifier is a must. Online classifiers are able to update the model in a sequential way, so the classifier, if needed, can constantly learn form new data, improving the quality of label proposal process. In order to optimize the learning process, the data are sorted according to the previously described criteria. A kernel-based online Perceptron classifier [16] is used because of its simplicity and efficiency. As previously mentioned, the main information used to detect non-informative frames is the color. In order to reduce the dimensionality of the data, each image represented as 24 million colors is quantized into 256 colors. The quantization is done using the training set of WCE images, where RGB space is clustered into 256 clusters using k-means [17] approach, and each centroid represents a point in the new colormap. As a result, each frame is represented by 256 color histogram. The score, for a given sample, takes this form:

$$S(x) = \sum_{j=1}^{K} \alpha_j K(v_i, x) \quad (3)$$

where $\langle(v_1, \alpha_1), ..., (v_k, \alpha_k)\rangle$ are the set of training vectors with the corresponding real estimated weights $(\alpha_1, ..., \alpha_k)$ by the learning algorithm when minimizing the cumulative hinge-loss suffered over a sequence of examples and $K()$ is a kernel function (in our case, we apply Radial Basis Function).

V. Results

For our experiments, we considered a set of 40 videos obtained using the WCE device. 10 videos were used to built the initial classification model M_1, and the other 10 to evaluate the proposed interactive labeling system. In the test, the 10 videos were sequentially processed. If needed, at each iteration, the training set could be increased by a new set of frames that improves the data representation. Additionally the validation set of 20 videos were used in order to evaluate the error of the final informative/non-informative frames classifier.

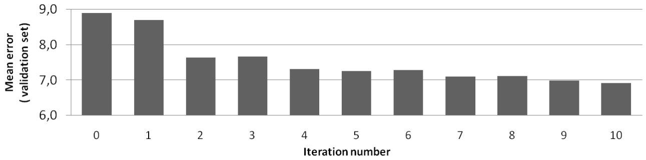

Figure 7. Mean error on validation set of 20 WCE videos.

In the experiments we show that: 1) the proposed system reduces the effort needed for data labeling, 2) the first criterion *Criterion 1 (C1): Distance of data to the classifier boundary and training data* gives the best results, 3) the global performance of informative/non-informative frames classifier is improving while enlarging training set, and 4) the LSH optimizes the computation process.Table 1 shows that all three proposed schemes reduce the number of clicks. Even random order improves a lot with respect to the naive approach. This phenomenon can be explained by the fact that the frame color in a certain video are well correlated.

From the results it can be concluded that the criterion 1 looks to be the best sorting criterion for interactive labeling. Intuitively, the samples that are far from the boundary are classified with high confidence. However, when dealing with the frames that are not similar to the ones in the training set (there are far from the one in the training set), the confidence gives uncertainty measure. Therefore, when introducing examples, where the classifier performs an error (user switches the label) to the model it is highly probable that the boundary will change using small amount of data. Introducing new data into the training set improves the final classifier performance and reduces the error by 2% after 10 iterations of the algorithm (where each iteration is new video introduced in the training set) (Fig. 7).

Furthermore, the LSH in average reduces the number of frames to query by more than 80%. This means that tested videos have about 20% of the frames that are ”strange”. While inserting new frames into the classifier model, at each iteration some data that are not represented in the training set are being found. The conclusion that can be drawn is that in order to create a good training set for informative/non-informative frame classification, the number of 20 WCE videos is not enough.

VI. Conclusions

In this paper, we review active learning methods and illustrate their application to optimize the process of labeling

TABLE I
RESULTS

Video	#frames	#strange fr.	#clicks Criterion_1	Criterion_2	Criterion_3
Video1	35847	4687	103	103	147
Video2	51906	10145	211	213	316
Video3	52777	5771	270	270	376
Video4	56423	13022	86	90	151
Video5	55156	7599	68	68	131
Video6	33590	17160	381	389	617
Video7	17141	1072	8	8	39
Video8	26661	5437	88	97	151
Video9	14767	1006	28	28	76
Video10	22740	1993	63	63	110
Mean clicks	-	-	1.5%	1.5%	2.9%

large amount of video frames. The system minimizes the user effort during the process of constructing a good representation of the WCE data is presented. The methodology is based on: 1) the detection of frames that enrich the training set representation, and 2) the interactive labeling system that allows to reduce the user effort, in the labeling process, using an online classifier which sequentially learns and improves the model. Three different sorting polices are defined and evaluated for the online classification: 1) *Distance of data to the classier boundary and training data*, 2) *Data density*, and 3) *Random order*. It is shown that by using adapted sorting criteria for the data, we can improve the label proposal process and in this way reduce the expert efforts. Finally, enlarging the training set with the non represented frames from new videos, we achieve an improvement of classification performance.

VII. Acknowledgements

This work was partly supported by projects: TIN2009-14404-C02, CONSOLIDER 2007-00018, and La Marato de TV3 082131.

References

[1] Iddan, G. and Meron, G. Et.al., ”Wireless capsule endoscopy”, Nature, 2000, 405, pp. 4-7.

[2] Given Imaging, Ltd., http://www.givenimaging.com/, 2000.

[3] Fireman, Z. and Mahanja, E., Broide E., Shapiro M., Fich L., Sternberg A., Kopelman Y. and Scapa E., ”Diagnosing small bowel Crohns disease with wireless capsule endoscopy”, Gut, 2003, Vol. 52, pp. 390-392.

[4] Schulmann K., Hollerbach S., Kraus K., Willert J., Vogel T., Mslein G., Pox C., Reiser M., Reinacher-Schick A. and Schmiegel W., ”Feasibility and Diagnostic Utility of Video Capsule Endoscopy for the Detection of Small Bowel Polyps in Patients with Hereditary Polyposis Syndromes”, The American Journal of Gastroenterology, 2005, 100 (1), pp. 27.

[5] Igual, L., Vitria J., Vilarino F., Segui S., Malagelada C., Azpiroz F. and Radeva P., ”Automatic Discrimination of Duodenum in Wireless Capsule Video Endoscopy”, IFMBE Proceedings, 22, 2008, pp. 1536-1539.

[6] Vilarino, F., Spyridonos, P. , DeIorio F., Vitria, J., Azpiroz F. and Radeva, P., ”Intestinal motility assessment with video capsule endoscopy: Automatic annotation of phasic intestinal contractions, IEEE TMI 29(2), pp. 246–259, 2010.

[7] Kang, J., Doraiswami, R.: ”Real-time image processing system for endoscopic applications”, IEEE CCECE03. vol. 3, pp. 1469-1472, 2003.

[8] Jung, Y., Kim Y., Lee D. and Kim J., ”Active blood detection in a high resolution capsule endoscopy using color spectrum transformation”, ICBEI. pp. 859-862, 2008.

[9] Coimbra, M., Cunha, J.: ”MPEG-7 visual descriptors: Contributions for automated feature extraction in capsule endoscopy”, IEEE TCSVT 16(5), pp. 628-637, 2006.

[10] Hai, V., Echigo T., Sagawa R., Yagi K., Shiba M., Higuchi K., Arakawa T. and Yagi Y., ”Adaptive Control of Video Display for Diagnostic Assistance by Analysis of Capsule Endoscopic Images”, Proceedings of the ICPR’06, III, 2006, pp. 980-983.

[11] Malagelada C., De Iorio F., Azpiroz F., Accarino A., Segui S., Radeva P. and Malagelada J.R.: ”New insight into intestinal motor function via noninvasive endoluminal image analysis”, Gastroenterology 135(4), pp. 1155–1162, 2008.

[12] Bashar, M., Kitasakaa T., Suenagaa Y., Mekadaa Y., and Morib K., ”Automatic detection of informative frames from wireless capsule endoscopy images”. Medical Image Analysis 14(3), pp. 449–470, 2010.

[13] Settles B., ”Active Learning Literature Survey, University of Wisconsin-Madison, No. 1648, Computer Sciences Tech. Report, 2009, pp. 246–259.

[14] Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: ”Gaussian processes for object categorization”. IJCV, 88, pp. 169-188, June 2010.

[15] Gionis A., Indyk P. and Motwani R., ”Similarity Search in High Dimensions via Hashing, Proceedings of the 25th ICVLDB, Series VLDB ’99, ISBN 1-55860-615-7, pp. 518–529, 1999.

[16] Orabona, F., Keshet, J., Caputo, B.: ”The projectron: a bounded kernel-based perceptron”, Procedings of the 25th ICML, pp. 720-727, 2008.

[17] Macqueen, J. B., ”Some Methods for classification and analysis of multivariate observations, Proc. 5th BSMSP, pp. 281–297, Vol. 1, 1967.

[18] Cohn D., Ghahramani Z., and Jordan M.I., ”Active learning with statistical models”, J. Artificial Intelligence Research, 4, pp. 129-145, 1996.

[19] Dagan I. and Engelson S., ”Committee-based sampling for training probabilistic classifiers”, Proceedings ICML, pp. 150-157, 1995.

[20] Fujii A., Tokunaga T., Inui K., and Tanaka H., ”Selective sampling for example-based word sense disambiguation”, Computational Linguistics, 24(4), pp. 573-597, 1998.

[21] Yu H., ”SVM selective sampling for ranking with application to data retrieval”, Proceedings KDD, pp. 354-363, ACM Press, 2005.

[22] Mitchell T., ”Generalization as search”, Artificial Intelligence, 18:203-226, 1982.

[23] Zhang C. and Chen T., ”An active learning framework for content based information retrieval”, IEEE Trans. on Multimedia, 4(2), pp. 260-268, 2002.

[24] Tur G., Hakkani-Tur D., and Schapire R.E., ”Combining active and semi- supervised learning for spoken language understanding”, Speech Communication, 45(2), pp. 171-186, 2005.

[25] Hauptmann A., Lin W., Yan R., Yang J., and Chen M.Y., ”Extreme video retrieval: joint maximization of human and computer performance”, Proceedings ACM Multimedia Image Retrieval, pp. 385-394, 2006.

[26] Settles B. and Craven M., ”An analysis of active learning strategies for sequence labeling tasks”, Proceedings EMNLP, pp. 1069-1078, 2008.

[27] Scheffer T., Decomain C., and Wrobel S., ”Active Hidden Markov models for information extraction”, Proc. CAIDA, pp. 309-318, 2001.

[28] Seung H.S., Opper M., and Sompolinsky H., ”Query by committee”, Proceedings ACM Computational Learning Theory, pp. 287-294, 1992.

[29] Tong S. and Koller D., ”Support vector machine active learning with applications to text classification”, Proc. ICML, pp. 999-1006, 2000.

[30] Melville P., Yang S.M., Saar-Tsechansky M., and Mooney R., ”Active learning for probability estimation using Jensen-Shannon divergence” Proceedings ECML, pp. 268-279, Springer, 2005.

[31] Goodman J., ”Exponential priors for maximum entropy models”, Proceedings HLT-NAACL, pp. 305-312. ACL Press, 2004.

[32] Guo Y. and Greiner R., ”Optimistic active learning using mutual information”, Proceedings IJCAI, pp. 823-829. AAAI Press, 2007.

[33] Roy N. and McCallum A., ”Toward optimal active learning through sampling estimation of error reduction”, Proceedings ICML, pp. 441-448. Morgan Kaufmann, 2001.

[34] Zhu X., Lafferty J., and Ghahramani Z., ”Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions”, Proceedings ICML W. Continuum from Labeled to Unlabeled Data, pp. 58-65, 2003.

[35] Geman S., Bienenstock E., and Doursat R., ”Neural networks and the bias/variance dilemma”, Neural Computation, 4, pp. 1-58, 1992.

[36] Cover T.M. and Thomas J.A., ”Elements of Information Theory”, Wiley, 2006.

[37] Hoi S.C.H., Jin R., and Lyu M.R., ”Large-scale text categorization by batch mode active learning”, Proceedings ICWWW, pp. 633-642, 2006.

[38] Dasgupta S. and Hsu D., ”Hierarchical sampling for active learning”, ICML, 2008.

[39] Dasgupta S., ”Two faces of active learning”, Theoretical Computer Science 412(19), pp. 1767–1781, 2011.

[40] Xu Z., Akella R., and Zhang Y., ”Incorporating diversity and density in active learning for relevance feedback”, Proc. ECIR, pp. 246-257. 2007.