Image Sequence Recognition with Active Learning Using Uncertainty Sampling

Masatoshi Minakawa, Bisser Raytchev, Toru Tamaki and Kazufumi Kaneda

Abstract—In this paper we consider the case when huge datasets need to be labeled efficiently for learning. It is assumed that the data can be naturally organized into many small groups, called chunklets, each one of which contains data from the same class, and many chunklets are available from each class. Each chunklet exhibits some of the typical variation representative for the class. We investigate how active learning methods based on uncertainty sampling perform in this setting, and whether any gains can be expected in comparison with random sampling. We also propose a novel strategy for selecting which chunklets to be selected for labeling. Experiments with face sequences containing variation in pose, expression and illumination conditions illustrate the proposed method.

I. INTRODUCTION

In pattern recognition, still the most predominant learning paradigm is based on supervised learning, where training data sets are gathered in advance and humans provide class/category information (labels) to be used during the training process. With the advent of the Big Data era, where cameras, microphones, RFID readers and ever more varied types of mobile/ubiquitous sensor devices and networks are gathering information, often 24 hours a day, the labeling process itself might need to be reconsidered: obviously labeling such huge amounts of data is becoming impractical and sometimes even impossible. Also, even with less-than-exabyte datasets, sometimes the labeling requires the costly time and effort of busy experts (e.g., medical doctors labeling images of different categories of tumors), and choosing which data samples to label in an optimal way becomes a necessity.

Fortunately, often the information content in such datasets is heavily redundant, which makes it possible to drastically reduce the number of required training labels by using active learning algorithms [1], [2], that choose selectively which samples (known as queries) should be labeled. In this paper we consider the specific case when the data is organized in groups or sets of samples of the same class/category. For example, consider the following cases: a surveillance camera is monitoring who is entering into a building, or a robot is tracking the faces of humans, or trying to learn different categories of objects by observing them from different views. In all these cases it would be much more efficient to consider the sequences of images of the same face (or the same object) taken during a short time interval as the smallest unit of data, rather than considering each image separately as such. In this way, the variability due to changes in factors which are not directly relevant to category information (like changes in illumination or view/pose in the examples above) can be spread within the sequence, allowing the classifier to concentrate on the variability between the categories. The benefits (in terms of increased recognition rates, robustness, etc.) stemming from this strategy have been consistently confirmed by research in image recognition [3]–[7] and more generally in machine learning [8], [9].

In this paper our aim is to investigate whether active learning methods still provide benefits, in comparison with random sampling, when the data is organized in groups or sequences, as explained above. It would be interesting to find out how the gains in increased recognition rates and efficiency following from the better use of information through the additional constraints inherent in the grouping organization would affect active learning methods.

II. ACTIVE LEARNING WITH IMAGE SEQUENCES

Here we illustrate the problem we deal with, using as a concrete example the case when one needs to perform face recognition from image sequences of different people. Such sequences can be easily obtained from surveillance cameras or monitoring cameras in conference rooms, etc., where it would be possible to track the faces of different individuals, so that it can be guaranteed that an image sequence obtained from a single track contains only face images from a single person. We call such sequences “chunklets” (following [8]), and a single chunklet would typically contain the object of interest (a face here) represented under a variety of poses, illumination conditions and facial expressions. Fig. 1 illustrates the concept, showing two face image sequences from two different people, one changing in pose, the other in facial expression (and both having changes in illumination to some degree). In Fig. 1 (a) all face images are treated as separate samples, which renders the problem quite difficult, as for example faces from different classes (people) but with similar pose would be nearer in feature space than faces from the same person but with different pose or expression. Once the faces are organized in chunklets, as Fig. 1 (b) shows, the problem becomes much easier due to the constraint that all images within a chunklet belong to a single class. Considering the class labeling process, the organization of the data in chunklets is also very beneficial, since now labeling a single face would automatically determine the labels of all other faces within the same sequence, due to the chunklet constraint. (Note that in Fig. 1 only one sequence per class is shown for illustration, but typically there are many unlabeled sequences from each class, both in
the training and test data sets.) Therefore, in the latter case training would be much faster and also we would expect that the number of labels required in an active learning setting would be greatly reduced.

This raises the following questions. First, considering active learning, what would be the gain in terms of reduction of the number of required labels when the chunklet constraint is used, compared to the case when each face sample is labeled independently. Next, when utilizing the chunklet constraint which is expected to significantly speed-up the learning process, would active learning still make sense, compared with random sampling, where the queries to be submitted for labeling are chosen randomly, i.e. without following any particular strategy as in the case of active learning. And finally, what strategy to use for chunklet labeling, as opposed to single sample labeling as in the standard active learning setting. In the rest of the paper, we will try to answer these questions, using as an example the aforementioned face-sequence recognition problem setting.

A. Uncertainty Sampling

Here we consider the general setting where a large subset \( \mathcal{U} \) of the available data is unlabeled (i.e. its class labels are unknown), while the class labels for only a small subset \( \mathcal{L} \) of the data are known. This is the standard pool-based sampling strategy for active learning [2], [10], where the samples in \( \mathcal{L} \) are first used to train a classifier (or an ensemble of classifiers), and then the samples in \( \mathcal{U} \) are evaluated according to some pre-defined measure of informativeness to select which among them should be submitted as a query for labeling. After that, \( \mathcal{L} \) is expanded to include the data corresponding to the newly obtained label, the classifier is re-trained and this process is repeated. Although many different strategies for selecting queries in an optimal way have been proposed in the literature (see [2] for a survey), here we limit our attention to the so-called uncertainty sampling framework [10], [11], in which an active learner selects as queries the instances about which it is least certain how to label. This is one of the most commonly used query-selection frameworks, which due to its low computational cost is suitable also for large datasets, and notwithstanding its simplicity it usually leads to significant gains in reduction of the number of necessary labels compared to random sampling [12]. However, note that it would be straightforward to extend the proposed chunklet-based labeling methods to other query-selection frameworks – all that is needed is to provide as an input some measure of informativeness for each sample, and the uncertainty sampling can be easily substituted by alternative measures of informativeness, if this is found advantageous for some concrete application.

Here we will first briefly review three of the most popular criterions for uncertainty sampling, which will be used and compared in the experiments described in section III. In each case it is assumed that a classifier has been trained on the available training data in \( \mathcal{L} \), and after the training the classifier can provide the posterior probability \( P_\theta(y|x) \) under model \( \theta \) that data \( x \in \mathcal{U} \) comes from class \( y_i \), where \( y_i \) ranges over all possible class labels.

The Least Confident criterion

: In the Least Confident criterion first the uncertainty of each sample \( x \in \mathcal{U} \) is calculated by

\[
UC_{LC}(x) = 1 - P_\theta(\hat{y}|x),
\]

where

\[
\hat{y} = \arg \max_y P_\theta(y|x)
\]

is the class label with the maximum posterior probability given data \( x \), and then the following sample \( x^* \) is selected for labeling:

\[
x^* = \arg \max_x UC_{LC}(x) = \arg \max_x (1 - P_\theta(\hat{y}|x)),
\]

that is, the data in whose prediction the model is least confident will be selected as the next query.
**The Margin Sampling criterion**

The Least Confident criterion considers only the most probable class label, ignoring information about the distribution of the other class labels. The Margin Sampling criterion [13] was proposed to improve on that by including also the posterior of the second most probable label. It maximizes

\[ UC_M(x) = 1 - (P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)), \]

where \( \hat{y}_1 \) and \( \hat{y}_2 \) are the first and second most probable class labels, respectively. By choosing for queries samples for which the margin between the first and second most probable class labels is small, the Margin Sampling criterion focuses on the most ambiguous data samples (while large margins imply that the classifier has less doubt in discriminating between the two most likely classes).

**The Entropy criterion**

For datasets with a large number of classes, the Margin Sampling strategy still ignores much of the information about the class probability distributions. As its name implies, the Entropy criterion uses the entropy [14], [15] to measure the uncertainty of a data sample

\[ UC_H(x) = -\sum_{i=1}^{K} P_\theta(y_i|x)\log P_\theta(y_i|x), \]

where \( K \) is the number of different classes, i.e. the Entropy criterion measures uncertainty by including information about the whole class probability distribution.

**B. Uncertainty Sampling with Image Sequences**

Here we explain how the query selection procedure can be modified when the data is organized into chunklets (groups of data of the same class/category). As illustrated in Fig. 2, each chunklet is a set of data samples (face images in this case) all belonging to the same class (same person), but representing some of the variability (face expressions) typical for the class – different chunklets would potentially represent different levels and content of this variability. However, once a classifier is trained using the data in the currently available labeled training set \( \mathcal{L} \), the posterior class probability distributions \( P_\theta(y|x) \) for each data \( x \in \mathcal{U} \) (the unlabeled dataset) can be calculated. Hence, the class probability distributions for all data samples inside each chunklet are available, as shown in the center of Fig. 2 (where A, B, C indicate different class labels), and from those, the uncertainty of each sample can also be calculated, using for example some of the uncertainty criteria from the previous subsection. The problem now becomes how to determine a measure of informativeness (uncertainty in this case) for the whole chunklet (i.e. how to determine the chunklet uncertainty shown on the right in Fig. 2). This is the main subject of this paper.

One possible way, which straightforwardly extends the standard uncertainty sampling procedure in the case when each data sample is considered separately to the chunklet-representation case would be to compare the uncertainty values of all samples within the chunklet and use the largest among them to represent the uncertainty of the whole chunklet. This can be calculated for all chunklets in \( \mathcal{U} \), and then the chunklet with the largest uncertainty can be selected as the next query to be labeled (i.e. all data inside the chunklet would be labeled). This can be expressed by the following equation:

\[ e^* = \arg\max_{c \in C} \ UC_A(\arg\max_{x \in c} UC_A(x)), \]
where $C$ is the set of all chunklets in $U$, $c$ are all chunklets in $C$, $c^*$ is the chunklet selected for labeling, and $UC_A$ is any of the uncertainty sampling criteria from the previous section (or any other measure of informativeness suitable for the problem). We will call this chunklet-query-selection procedure, the max-max method.

However, selecting the sample with the maximum uncertainty to represent the whole chunklet might be a wasteful strategy, as it could throw away information available in the chunklet. For example, in the face recognition case a certain face image might represent the subject in a pose or with a facial expression which is difficult to classify at the present stage of the learning process (especially in the early stages of the learning), or simply the data might be an outlier (due to some failure in the pre-processing process, etc.) which is not typical for the whole chunklet. At the same time, in the chunklet might be present some data (e.g. a face in frontal position and with neutral expression) which the classifier might be able to classify with a high degree of confidence. Therefore, one could argue that it is actually the least uncertain sample which has to be chosen to represent the chunklet. We call this strategy the min-max method and it can be written as

$$c^* = \arg\max_{c \in C} UC_A(\arg\min_{x \in c} UC_A(x)).$$

(7)

Still, one could argue that both the max-max and the min-max strategies represent two extreme views, and using some statistical measure which considers the uncertainties of all data samples within the chunklet, like the average or median, etc., might be a better representative for the uncertainty of the whole chunklet. In the experiments described in the following section, we also used a hybrid method which represents a compromise between the max-max and the min-max methods. The hybrid method alternates at each step between the max-max and the min-max strategies, i.e. at one step the max-max methods is used and at the next step the min-max step is used. This would allow it to compensate somewhat if one or both of the extreme views might be wrong at certain stages of the learning process. We performed experiments also using the median of the uncertainties of all data samples within a chunklet, but this seemed to perform very similar to the hybrid method (except in the Margin sampling case, where it performed worse than min-max, max-max or the hybrid method – still better than random sampling though), so we don’t show results for the median in order not to overcrowd the graphs in the figures.

III. EXPERIMENTS

We have performed several experiments to illustrate the behavior of the uncertainty sampling-based active learning algorithms described above in the case when the data can be organized in chunklets, here shown in the context of a face recognition task. For our experiments we used the publicly available MOBIO database [16]. This data set contains a large number of face video sequences acquired primarily on a mobile phone (NOKIA N93i) between August 2008 and July 2010 in six different sites from five different countries. The videos were recorded in 6 different sessions under different environmental conditions. In each session, the participants were recorded while being asked to answer a set of 21 different questions, i.e. 21 videos per session per subject are available. The questions were of different type, including both free speech and set speech. As a result of this experimental setting the video sequences contain natural facial expressions, changes in facial pose and illumination conditions, as those accompanying natural human communication.

We used a subset containing the data for 30 subjects (15 men and 15 women) selected randomly, resulting in a data set containing about 90000 faces. The faces were extracted from the original video sequences using an eye detector available from the OpenCV library [17]. The resulting face images were resized to $60 \times 60$ pixels and the feature dimension was further reduced by PCA to about 50 dimensions (keeping 90% of the variance). As a classifier we used an implementation of Support Vector Machines which can provide posterior class probabilities, as implemented in the LIBSVM library [18]. The recognition rates reported below are the result of using a 6-fold cross validation, where in turn data from 5 sessions was used for training and the remaining one session was used as a test set. As initially we assume none of the data in the training set is labeled ($L$ is initially empty), we performed k-means clustering [19] using $k = 30$ clusters, and labeled the faces from each class which were nearest to the cluster centers found by the k-means clustering procedure.

Next, we compare the results obtained when active learning using the uncertainty criteria from the previous section were used in combination with sequence-based chunklet-query selection or individual face-query selection. These were also compared to random sampling (no use of active learning strategies) and full supervised learning (where all data in the training set is labeled, i.e. all data is in $L$ and no data is in $U$).

![Fig. 3. Experimental results when the Entropy criterion for uncertainty sampling was used](image)

Fig.3 shows the results obtained when using the entropy criterion. The red line at the top (supervised) shows the
average recognition rate when fully supervised learning was used, using the information from all labels in the training sets. The aim of active learning is to achieve a similar rate, however using only a small fraction of the labels (the number of labels used in the active learning or random sampling procedures are given on the horizontal axis). Somewhat surprisingly, we see from the figure that if active learning is used only with the individual faces, without using the chunklet constraint (the gray curve denoted as “no-chunk AL”), it performs much worse than random sampling which also does not use the chunklet-constraint (the green curve denoted “no-chunk random”). The situation reverses, however, when the chunklet-constraint is used. Now the proposed min-max method performs much better than random sampling with the chunklet-constraint (denoted as “chunk-random” in the figure) and reaches (and even surpasses) the full supervised learning method at around 200 labels (which is less than 10 labels per face class). The max-max method performs well in the initial stages of the learning process, however it is surpassed by the chunklet-based random sampling at the later stages of the learning process. The hybrid method which is a compromise between the min-max and max-max performs better than random sampling.

Fig. 4 shows the results obtained when using the Least Confident criterion for entropy case (however now it takes a little bit longer for the min-max to reach the fully supervised case).

The results for the Margin sampling criterion are shown in Fig. 5. Now the situation is a little bit different compared with the previous cases. Now active learning performs always better than random sampling, even in the case when the chunklet-constraint is not used, and active learning without the chunklet-constraint is even better than random sampling using the chunklet-constraint. Still, the chunklet-based active learning methods perform best, however now in the early stages of the learning process the max-max criterion performs better than the min-max (which catches up soon), and the hybrid method performs best.

IV. CONCLUSIONS

In this paper we investigated the problem how active learning methods based on uncertainty sampling behave in the case when the data is organized in groups or sequences (called chunklets in the paper) representing different forms of variability typical for the class/category to which the data group belongs.

We showed that if the chunklet constraints are not utilized it is possible that active learning can actually perform inferior to random sampling. We also proposed several criteria which by using the information available in the chunklets were able to outperform random sampling in all cases, and reach similar results to a fully supervised learning, however needing only a small fraction of the data to be labeled.

In a future work, it would be interesting to consider investigating alternative definitions of chunklet uncertainty, which could utilize more fully the information available in the posterior class probability distributions of the data inside the chunklets. Also, it is necessary to investigate how the chunklet-constraint would affect other forms of active learning, based on alternative criteria other than the uncertainty sampling framework used in the present work, and also to validate the proposed methods on other datasets.

REFERENCES


