Pattern Recognition Letters 73 (2016) 76–82

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Active graph based semi-supervised learning using image matching: Application to handwritten digit recognition☆

Hubert Cecotti*

Faculty of Computing and Engineering, Ulster University, Magee campus, Londonderry, BT48 7JL, Northern Ireland, UK

ARTICLE INFO

Article history:
Received 20 July 2015
Available online 3 February 2016

Keywords:
Active learning
Semi-supervised learning
Character recognition
Image matching

ABSTRACT

With the availability of large amounts of documents and multimedia content to be classified, the creation of new databases with labeled examples is an expensive task. Efficient supervised classifiers often require large training databases that are not always immediately available. Active learning approaches solve this issue by querying an expert to set a label to particular instances. In this paper, we present a novel active learning strategy for the classification of handwritten digits. The proposed method is based on a k-nearest neighbor graph obtained with an image deformation model, which takes into account local deformations. During the active learning procedure, the user is first asked to label the vertices with the highest number of neighbors. Thus, the expert sets the label to the examples that are more likely to propagate their labels to a high number of close neighbors. Then, a label propagation function is performed to automatically label the examples. The procedure is repeated until all the images are labeled. We evaluate the performance of the method on four databases corresponding to different scripts (Latin, Bangla, Devnagari, and Oriya). We show that it is possible to label only 332 images in the MNIST training database to obtain an accuracy of 98.54% on this same database (60000 images). The robustness of the method is highlighted by the performance of handwritten digit recognition in different scripts.

1. Introduction

The fast increase of new documents and multimedia content to be classified is a source of new challenges in pattern recognition and machine learning [13]. Documents and multimedia data are exponentially growing thanks to the internet and the development of portable devices that can acquire images. In addition, cultural heritage collections are being digitized, and made available through online tools. New automatic methods must be provided to both automatically index and search through the documents because not enough manpower is available to provide useful annotations on the large volume of digitized documents [32]. With the emergence of the Big data paradigm and the creation of new classification problems, pure supervised techniques may not be able to cope with the fast increase of classification tasks, which possess only few labeled examples. Because the requirement of a classification task can evolve rapidly over time, it is essential to propose a fast evaluation of the potential performance. This diagnostic may infer a different type of approach in a later stage. In multiclass classification tasks, several approaches are possible depending on the type of data. First, the training data is well identified, and a ground truth is available, therefore classical supervised classification techniques can be used. In such a passive supervised learning, the goal of a learner is to infer an accurate predictor from labeled training data. The expert is passive as the system uses only existing labeled training examples. The labeled training data are input-output pairs (x, y): the feature set x describing the example, and its corresponding label y. Second, image retrieval methods can be used when the number of examples is too small, and when there is a large variability across examples [1,16]. Third, the images belong to a new type of classification problem, and an efficient technique has to be provided to facilitate data labeling, i.e., the creation of the ground truth. The creation of a ground truth is in fact an important aspect because providing accurate labels can be a challenging task that requires the full attention of the users. This task can require several users to validate the results. In the case of medical images, the ground truth can only be created by an expert. Therefore, the ground truth estimation, as a major component of a pattern recognition system, can be time consuming and costly.

In active learning, each example of the training database is initially unlabeled. However, the active learner is allowed to request the ground truth, the label y, of any particular example x in the training database [35]. The requests can be made after a

☆ This paper has been recommended for acceptance by Ajay Kumar.

* Tel.: +44 2871675276.
E-mail address: hub20xx@hotmail.com, h.cecotti@ulster.ac.uk

http://dx.doi.org/10.1016/j.patrec.2016.01.016

non-supervised learning technique (e.g. examples are clustered, and the centroids are then labeled by an expert), or online (sequentially) in order to adapt the classifier to previous label requests. The objective of these methods is to discover the labels of examples while minimizing the number of manually labeled examples [43]. This solution is particularly adapted in large data collections where there exists a strong disparity between the availability of labeled and unlabeled data. In such a case, a challenging task is to provide semi-automatic, high accuracy labeling mechanisms. To some extent, active learning is an easier task than semi-supervised learning (SSL), because in SSL the labeled examples are predefined. The existing labeled examples may be outliers, and/or they may not provide good seeds for label propagation. In active learning, we can distinguish two strategies that require the use of an expert. First, the expert is needed because a potential confusion between two examples is detected, and this ambiguity should be raised by an expert. Second, the expert is needed to label the examples that have the highest potential to be beneficial in the learning procedure. The latter approach is considered in the proposed method: the expert sets the best seeds that can reliably propagate their labels to other unlabeled examples.

In this paper, we propose a new active learning method that combines an efficient distance measure based on Image Distortion Model Distance (IDMD) [20], a greedy SSL approach, and active learning. The efficient distance measure allows us to obtain a robust graph that respects the manifold assumption: if two examples are similar then their corresponding vertices in the graph are connected. The SSL part is dedicated to the label propagation to local neighborhoods. Finally, the active learning step guides the method to the most relevant examples to label. To show the relevance of the method, we use four databases of single handwritten digits. For the Latin script, the performance of single handwritten character recognition is typically sufficiently high when the number of images to train a classifier is high. The supervised learning methods include deep learning architecture such as convolutional neural networks [10], Support Vector Machines (SVM) [12], and their combination [22,29]. However, the accuracy of single handwritten character recognition remains below 100% in some scripts because documents are not properly conserved, and are therefore noisy once they are digitized. Furthermore, a large variability across writers, with several styles and different glyphs for a same digit or character, can become an obstacle. For all these reasons, it is essential to propose new methods to maximize the accuracy while minimizing the number of labeled training examples. Moreover, the recognition of some characters can be impossible without any contextual information and may only be achieved by a person. The remainder of the paper is organized as follows: First, we give an overview of image matching techniques in character recognition and active learning methods in Section 2. Then, we describe the new method in Section 3. In Section 4, we present the four handwritten databases. The active learning strategy is then evaluated in Section 5. Finally, the performance of the proposed approach is discussed in Section 6.

2. Related work

2.1. Image matching

Image matching techniques in large databases are usually not used due to the high processing time that is involved, e.g. the computation of distances between the test image and the prototypes. Elastic matching techniques can be classified into two categories: parametric and non-parametric [39]. It is typically seen as an optimization problem of two-dimensional warping (2DW). This problem is directly related to point matching, which has to deal with the existence of outliers and geometric transformations that may require high dimensional non-rigid mappings [14]. Deformations in handwritten characters can be of two types: first, the global or large deformations such as rotation (with limited angles), scaling, translation; and second, the local deformations that include changes of stroke direction, curvature, and length of the lines. The local deformations that are involved by the thickness of the character depend on the pen/pencil that is used. Due to the different types of deformations that can occur within the same character, it is difficult to determine generic models of deformations. An efficient distance for image classification that takes into account local deformations was proposed by Keysers et al. [20]. They determined that more complex models (*e.g.* 2-dimensional warping) do not necessarily represent better models compared to the simple image distortion model. We define the distance L_p between two images A and B of size $N_s \times N_s$ by:

$$L_p = \left(\sum_{i=1}^{N_s} \sum_{j=1}^{N_s} |A(i,j) - B(i,j)|^p \right)^{1/p} \tag{1}$$

When $p = 2$, it corresponds to the Euclidean distance. The image distortion model distance (IDMD) takes as input two images A and B of size $N_s \times N_s$ that can have w_2 multiple channels. In this study, each channel corresponds to the graylevel image processed with a convolution filter. The distance is then computed through a range of pixels from N_{min} to N_{max}. For each pixel (i_1, j_1) of A, a square image patch, centered at the pixel, is compared to a square patch of same size at the same region in the image B (channel wise, if there are multiple channels). To cope with local variations, the position of the square patches in the image B is allowed to be slightly shifted (within distance w_0). Thus, each patch in A is compared to multiple patches in B around the same pixel location, and the minimum computed distance is taken.

$$\mathrm{IDMD(A,B)} = \sum_{i_1=N_{min}}^{N_{max}} \sum_{j_1=N_{min}}^{N_{max}} d_1(i_1, j_1) \tag{2}$$

where

$$d_1(i_1, j_1) = \min_{(i_2, j_2) \in \{-w_0; w_0\}^2} d_2(i_1, j_1, i_2, j_2) \tag{3}$$

$$d_2(i_1, j_1, i_2, j_2) = \sum_{i_3=-w_1}^{w_1} \sum_{j_3=-w_1}^{w_1} \sum_{i_4=1}^{w_2} |v_1(i_4) - v_2(i_4)|^p \tag{4}$$

where $v_1(i_4)$ and $v_2(i_4)$ are the pixel values at the following coordinates in the $i_4 th$ channel of the image $(1 \leq i_4 \leq w_2)$:

$$v_1(i_4) = A_{i_4}(i_1 + i_3, j_1 + j_3) \tag{5}$$

$$v_2(i_4) = B_{i_4}(i_1 + i_2 + i_3, j_1 + j_2 + j_3) \tag{6}$$

For each pixel (i_1, j_1), a displacement field of size w_0 is used in each direction (it corresponds to the elements of a square window of size $2w_0 + 1$), a square window for the consideration of the neighborhood pixels of size $2w_1 + 1$, and the sum of w_2 values, which corresponds to w_2 filtered images. It is worth noting that the *min* function aims at including a relative invariance to local deformations. The image has to be placed in a larger image with a border of the background color to include the possible shifts in the four directions, hence N_{min} and N_{max} are set to $w_0 + w_1$ and $N_{min} + N_s$, to take into account the size of the filters and the displacement fields. Finally, the Euclidean distance (L_2) between I_1 and I_2 is similar to IDMD with the following parameters: $w_0 = 0$, $w_1 = 0$, $w_2 = 1$, $p = 2$.

2.2. Semi-supervised learning and active learning

In semi-supervised learning, several approaches have been proposed [44]. Transductive SVMs optimize margins of both labeled