Generating Synthetic Data to Allow Learning from a Single Exemplar per Class

Liudmila Ulanova, Yuan Hao, and Eamonn Keogh

Department of Computer Science & Engineering, University of California, Riverside, USA
{lulan001,yhao,eamonn}@cs.ucr.edu

Abstract. Recent years have seen an explosion in the volume of historical documents placed online. The individuality of fonts combined with the degradation suffered by century old manuscripts means that Optical Character Recognition Systems do not work well here. As human transcription is prohibitively expensive, recent efforts focused on human/computer cooperative transcription: a human annotates a small fraction of a text to provide labeled data for recognition algorithms. Such a system naturally begs the question of how much data must the human label? In this work we show that we can do well even if the human labels only a single instance from each class. We achieve this good result using two novel observations: we can leverage off a recently introduced parameter-free distance measure, improving it by taking into account the "complexity" of the glyphs being compared; we can estimate this complexity using synthetic but plausible instances made from the single training instance. We demonstrate the utility of our observations on diverse historical manuscripts.

Keywords: Classification, Semi-Supervised Learning, Historical Manuscript, Handwriting Analysis.

1 Introduction

The classification of individual glyphs is typically the first step in historical document processing. The variety of texts (hundreds of languages, tens of thousands of handwriting styles/handmade fonts), combined with the degradation often suffered by century old manuscripts, precludes the adoption of a "one-size-fits-all" off-the-shelf Optical Character Recognition (OCR) Systems.

Most Semi-Supervised Learning (SSL) techniques make explicit assumptions which are violated or only partly true in our domain of interest [3]. In particular, the *smoothness assumption* can be violated in a special way that does not seem to be well appreciated. Recall that it requires that "*(objects) which are close to each other are more likely to share a label*" [3]. However, this assumption can be violated in an unexpected way: "complex" objects tend to be closer to other objects that are "simple," at least under some distance measures such as the recently introduced CK-1 distance [2]. In Fig. 1 we show a clustering that hints at this [5, 11]. This "complexity bias" violates the notion that objects that are close to each other are more likely to share a label, at least for some classes.

A.J. Machado Traina et al. (Eds.): SISAP 2014, LNCS 8821, pp. 189–194, 2014.
DOI: 10.1007/978-3-319-11988-5_17

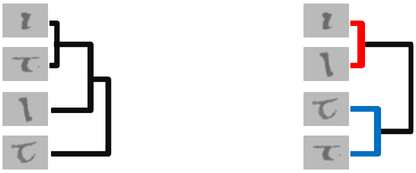

Fig. 1. (*left*) A clustering of examples from [5] suggests that the distance measure has difficulty clustering objects with different complexities, such as the simple ‘i’ and more complex ‘t’. (*right*) If we compensate for these differences in complexity we can do much better.

Given enough training data we can learn the amount of “complexity bias” for each class and compensate for it. However, this opens up a “chicken-and-egg” paradox, as we are using SSL to mitigate the *lack* of training data. As we shall show, we solve this problem by creating additional synthetic examples with a simple random distortion model. As hinted at in Fig. 2, we can easily produce plausible variations of hand-press or handwritten letters.

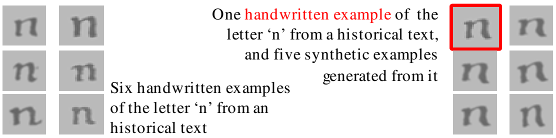
Six handwritten examples of the letter ‘n’ from an historical text
One handwritten example of the letter ‘n’ from a historical text, and five synthetic examples generated from it

Fig. 2. (*left*) Six examples of a handwritten letter. (*right*) We can take a single letter (red / highlighted example) and produce natural looking variations of it with a simple distortion model.

The rest of the paper is organized as follows. In Section 2 we discuss related work and background for our research. In Section 3 we introduce our proposed method. Section 4 presents experimental evaluations. Finally, Section 5 offers conclusions and a discussion of avenues for future research.

2 Related Work and Background

While there is a plethora of classification algorithms available, the simple Nearest Neighbor (NN) algorithm is known to be surprisingly competitive in many domains. This is because the algorithm can use *any* distance measure, including ones that can “carve out” decision boundaries that are not within the representation power of decision trees, etc. In this work we propose to leverage off a recently introduced distance measure called the CK-1 distance [2]. The CK-1 distance differs from other methods (Gabor filters, Fourier transforms, Markov random fields, wavelets, etc.) in two important ways. First, it considers shape *and* texture simultaneously. Second, it is completely parameter-free, freeing us from the need to obtain data to learn parameter settings, and greatly reducing the probability of overfitting (with no parameters to *fit*, one cannot over*fit*). The CK-1 is a compression based distance measure. The distance ranges between zero and “soft” one. If two objects are very similar to each other the distance is close to zero, whereas for very dissimilar objects the distance is close to one or slightly greater. Due to its simplicity and effectiveness we use CK-1 distance measure in this work; however, it is not a necessary condition for the utility of our ideas. CK-1 has proven its efficiency in historical documents processing domain [8], but in a bit different sense. While Hu et al. apply this distance measure to initial letters mining (intrinsically *textures*); we expand it to all glyphs (intrinsically *shapes*). However, there are two problems we must solve in order to use CK-1. The first is data scarcity. All classification algorithms benefit from more data; however, our explicit problem statement allows us to have as few as one exemplar per class. The second problem, which was hinted at in Fig. 1, is less well appreciated in the literature. At least some distance measures may overestimate the distance between "complex," but nevertheless similar, objects. For example, in our domain, letters such as $\mathscr{A}$ and $\mathscr{B}$ are complex, at least relative to the more prosaic versions, **A** and **B**. In this case the difference in complexity is related to particular typeface/handwriting. However, even *within* a single typeface, there are differences in complexity, ranging from the simple *single-stroke* letters such as **I** and **O**, to more complex *multi-stroke* letters such as **W** and **E**. The observation that differing complexities cause problems for nearest neighbor classification has been forcefully shown for time series classification [1]. Moreover, as we shall show below it is also the case for classifying glyphs with the CK-1 distance measure. Note that we are not making any claims with regard to other shape distance measures[1].

Synthetic data generation techniques are widely used to supplement datasets that do not have a sufficient number of instances for a given task [6]. If each exemplar can be described by a feature vector, then the problem of synthetic data generation can often be solved by a technique as simple as adding random Gaussian noise to copies of the original vectors, or by averaging randomly chosen vectors from the same class (i.e., SMOTE and its variants [4]). The problem becomes more complicated if we are dealing with objects that cannot be easily represented by feature vectors. In a recent work Yang et al. proposed a method of data densification in image domains [14]. Their insight is that they can forgo creating synthetic *exemplars*, and simply create synthetic *points* in the distance space. Such points make the estimation of the data manifold more accurate, and can thus improve retrieval accuracy.

While this work is closest in spirit to ours, we *do* need to create actual synthetic images in order to learn the potential biases of our distance measure, and correct them. To produce synthetic exemplars we apply transformations similar to those proposed by Ha et al. [7]. This model captures majority of variations in writing and produces plausible results shown in Fig. 2.

Wang et al. [13] introduced the Adjusted k-Nearest Neighbors Rule, considering "*influence region for each training example*." They constructed this region as a sphere centered on the example "*that is as large as possible without enclosing a training example of a different class*". After this they rescale the distances to each training sample as distance divided by the radius of the influence region. This approach is similar to ours, because it takes into account the "density" of training items in the

[1] Although preliminary work suggests that other distance measures also have difficulties in datasets with classes of varying complexity.

distance space. However, this approach requires parameters to be adjusted and optimized for each particular problem, something we are anxious to avoid.

We will provide necessary definitions before describing our algorithm.

Definition 1: A *labeled example* e_i of a class E_i is a *human* annotated glyph. Similarly, an *unlabeled example* is any example of the same glyphs not labeled by human.

After selecting e_i we can generate synthetic data based on e_i by a *distortion model*:

***Definition* 2**: A *distortion model M* is the method to modify labeled glyphs to generate synthetic data $\{S_{i,j}\} = M(e_i)$, where j denotes index of synthetic exemplars in a particular class i ($1 \leq j \leq$ const).

Note: our distortion model M is only one of many methods that can generate synthetic data. Further discussion of synthetic data generation techniques is beyond the scope of this paper; we refer the reader to [4] and [7], and therein.

To classify an unknown item we have to consider distance (similarity measure) between this item and labeled items from the training set. Our approach exploits the *correction* of distance calculated by some known algorithm.

Definition 3: The *corrected distance* between the query image q and *any* object $S_{i,j}$ in the training dataset of the i^{th} class is a distance under some distance measure (i.e., CK-1) divided by some correction factor μ_j.

3 Proposed Method

As the dendrogram shown in Fig. 1 (*left*) suggests, classification using the CK-1 distance measure sometimes does not correspond to the ground truth. We have observed that in cases where both shapes are "complex" (i.e., consisting of several "strokes" such as '**f**' or '**x**') the distance between them is greater than in cases where the shapes are "simple" (i.e., containing a single stroke such as '**l**' or '**o**'). Fig. 3 presents a visual intuition of this phenomenon. In terms of absolute distances the unknown object lies slightly closer to the nearest '**o**'. Intuitively however, we may feel it is likely to belong to the '**ft**' class because this class is sparse and the mean distance between two instances of this cluster is relatively larger than in '**o**' class.

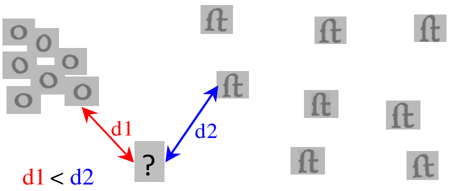

Fig. 3. While the unknown "**?**" object is slightly closer to the nearest o than to the nearest ft, we intuitively feel it is more likely to belong to the latter class (exemplars are from [10])

The classic nearest neighbor algorithm does not take into account the density of each cluster. In order to mitigate this shortcoming, we must correct for density, that can be characterized as the mean of intraclass distances. Our approach is inspired by the inverse-square law which is widely applied in physics. Let us consider an analogy to Newton's law of universal gravitation which follows an inverse-square law.

$$F = G\frac{mM}{r^2} \tag{1}$$

In (1) F is the force between two objects which masses are m and M, r is the distance between centers of mass of these objects and G is the gravitational constant. Analogously, in case of the classification problem we can consider the distance between the unknown object and the nearest neighbors of each class as r (denoted as r_i for each class i) and the mean of distance inside each class as M (for each class i denoted M_i). Since we simply need to compare the resulting values of F given by distance measuring between the unknown object and the nearest neighbor in each class, we are not interested in the values of G and m because they are the same in all cases. Thus, we need to compare these ratios: M_i/r_i^2 and M_j/r_j^2. Each ratio shows the "force" of attraction of the unknown object by each existing class. Therefore, the unknown object should be considered belonging to the class with the greatest "force." Recall that the nearest neighbor classifier assigns unknown objects the class label of a known object with the least distance value. Therefore, we can simply look for the least value between: $r_i/\sqrt{M_i}$ and $r_j/\sqrt{M_j}$ Thus, we can consider *division by square root* of the mean as the appropriate correction factor for the distance.

We initially imagined that creating synthetic data would be a major challenge. However, we found that simply applying tiny amounts of the affine transformations *homothety*, *rotation* and *shear mapping*, produced new images that are both visually very convincing (cf. Fig. 2) and closely modeled the true distributions of real data.

We present two different algorithms: supervised learning (SL) and semi-supervised learning (SSL). SL algorithm classifies items using only exemplars from the training set without addition of newly-classified items from the testing set to the training set. In contrast the SSL algorithm adds newly-classified instances to the training set and, therefore, performs next item classification using both training (generated) data and newly labeled instances from the testing set. For both algorithms we generate synthetic data randomly extracting one example from each class, and then distort this example applying the distortion model. After the training set is created we calculate distances between exemplars in one class with each other and find the mean of these distances to use it as our measure of class density (i.e. correction factor).

4 Experimental Evaluation

Table 1 shows the accuracy improvement of glyphs classification using distance correction over pure nearest neighbor approach. As we can see, our method demonstrated better performance than pure nearest neighbor approach.

Table 1. Classification accuracy (in percent) for the datasets: 1 – Chinese, 2 – G114 Verard Grosromain, 3 – R118 Garamount Grosromain, 4 – Liber Floridus, 5 – Petroglyphs

	Original accuracy, %					**Accuracy improvement, %**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
SL	96.6	81.6	95.8	91.6	81.8	1.3	4.3	1.3	2.5	16.4
SSL	98.0	86.0	85.9	97.1	81.6	0.5	5.3	1.4	1.2	13.1

We have built a webpage [12] to more extensively document the experiments in this paper to ensure reproducibility. We tested our SL and SSL algorithms using a single, randomly chosen instance from each class. In every case, we averaged our results over 100 random runs. We evaluated our approach with both European, Chinese handwritten historical documents and petroglyphs from 5 datasets [5, 9, 10, 15].

5 Conclusions and Future Work

We have shown a method that allows classification of glyphs using only one exemplar of each class, by exploiting synthetic data and correcting distance calculations for the complexity of the glyph shapes. Experimental evaluation on diverse datasets demonstrated significant improvements in accuracy. We have committed to keeping a webpage with all the code and data we used online for at least five years, so others can check/reproduce and build upon our work [12].

For future work we consider expanding our techniques to other areas of images recognition as well as exploiting different distance measures for comparison of images similarity.

References

1. Batista, G., Wang, X., Keogh, E.J.: A Complexity-Invariant Distance Measure for Time Series. In: Proc. of the SDM 2011, pp. 699–710 (2011)
2. Campana, B., Keogh, E.: A Compression Based Distance Measure for Texture. In: Proc. of the SDM 2010, pp. 850–861 (2010)
3. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning. MIT Press, Cambridge (2006)
4. Chawla, N., Bowyer, K., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
5. Derolez, A., Lamberti, S.: Audomari Canonici Liber Floridus, Codex Autographus Bibliothecae Universitatis Gandavensis, Ghent (1968)
6. Eno, J.: Generating Synthetic Data to Match Data Mining Patterns. IEEE Internet Computing 12(3), 78–82 (2008)
7. Ha, T., Bunke, H.: Off-line handwritten numeral recognition by perturbation method. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(5), 535–539 (1997)
8. Hu, B., Rakthanmanon, T., Campana, B., Mueen, A., Keogh, E.: Image Mining of Historical Manuscripts to Establish Provenance. In: Proc. of the SDM 2012, pp. 804–815 (2012)
9. Indiana MAS Project, `http://indianamas.disi.unige.it/`
10. PaRADIIT Project, `https://sites.google.com/site/paradiitproject/`
11. Roy, P., Rayar, F., Ramel, J.Y.: An efficient coarse-to-fine indexing technique for fast text retrieval in historical documents. In: DAS 2012, pp. 150–154 (March 2012)
12. Supporting web page, `https://sites.google.com/site/singleexemplar/`
13. Wang, J.-G., Neskovic, P., Cooper, L.N.: An adaptive nearest neighbor algorithm for classification. In: Proc. of ICMLC 2005, pp. 3069–3074 (2005)
14. Yang, X., Bai, X., Köknar-Tezel, S., Latecki, L.J.: Densifying Distance Spaces for Shape and Image Retrieval. Journal of Mathematical Imaging and Vision, 1–17 (2012)
15. Zhang, X., Nagy, G.: The CADAL calligraphic database. In: Proc. of the HIP 2011, pp. 37–42 (2011)