Document classification based on what is there and what should be there

Noga Levy[1], Lior Wolf[1*], and Peter A. Stokes[2†]

1 Blavatnik School of Computer Science, Tel Aviv University, Israel

2 Department of Digital Humanities, King's College London, UK

Introduction

Some of the key questions in paleography are those of classification, namely trying to ascertain when and where a given manuscript was written, and — if possible — by whom. Paleographers bring many skills and tools to bear on these questions in what is often a complicated and laborious task requiring reference to paleographic, linguistic and archaeological data, among others. Because it is difficult to quantify the degree of certainty in the final readings and assessments, or even to articulate the arguments underlying these readings, experts have begun to develop computer-based methods for paleographic research in which the description of the various findings is made explicit (Ciula, 2005; Stokes, 2008; Aussems and Brink, 2009; Hofmeister et al., 2009).

Some of these computer-based approaches involve little or no human intervention. However, others require manual selection of regions in the image or manual recording of descriptors, that is, of features in the handwriting which are considered significant (Ciula, 2005; Stokes, 2008). Evaluating the significance of the features can be improved using statistical analysis (Levy et al., 2012). Such manual selection raises a key challenge in any system of descriptors, namely that of attribute repeatability among documents of the same category. Would two different people necessarily record the same descriptors for a given sample of writing? Surely some significant features would then be overlooked? If so then what are the implications, both for the accuracy of the results and for the perceived "objectivity" of the method. A descriptor that is marked as existing in a document is likely to exist; however, a descriptor might be unmarked due to an omission or simply because it is not present in the part of the manuscript that is available for inspection. Moreover, even very discriminative descriptors (those which are very important for distinguishing date, location or scribe) might not be present where expected due to scribal variance within the same location and date.

In order to overcome this challenge, we suggest a new statistical tool that allows us to hypothesize which attributes should be turned on — in other words, which attributes are likely to have been omitted due to the limits of selection — and then to perform classification on the augmented data. Our results demonstrate that this tool is effective in computer-based document classification.

* LW was supported by a personal research grant from Google Inc.

† The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7) under grant agreement no. 263751.

Overview and results

A dataset consisting of scribal hands in English Vernacular minuscule, ca. 990 – ca. 1035, is used, where "scribal hand" here refers to a single stint or block of writing by one person (Stokes, 2005; Stokes et al., 2013). These samples are spread across some 198 manuscripts and range from the main text of the book to later additions and notes or glosses between the lines or in the margins; they therefore can include anything from hundreds of pages to just one or two words.

The hands were described using 289 descriptors (Stokes, 2008), where each descriptor indicates whether a certain letter-form is present; more precisely, whether a grapheme (or group of similar graphemes) written as specific allograph(s) appear in the manuscript, as well as forms of certain parts of letters such as ascenders, descenders, and pen-angle. Examples of augmented representations are presented in Fig.1. Every sample of handwriting is described by its known or predicted place of writing (where possible) and the estimated range of dates of writing. The date and localisation is based on external evidence wherever possible, or otherwise by an expert assessment of paleographical judgment (Stokes, 2005).

We focus on the samples whose place of writing is unknown but there is an educated guess to their origin, and try to verify their assumed place of writing. Overall, there are 67 such samples. The samples for which the place of writing is known, totaling 120 documents, serve as the training set. There are seven categories, such as Canterbury, Sherborne, and Worcester.

The baseline classifier we employ is the popular Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995) since it is known to be robust and to provide results that are often very close to the best obtainable, and because it outperformed other classifiers which we tried such as adaBoost (Freund and Schapire, 1995) and classification trees (Breiman et al., 1984). For each location-based category a model is learned by considering all documents which are known to belong to this location as the positive training set, and all other labeled documents as the negative training set.

Given a handwriting of an unknown origin, we apply all location-based models and compare the model producing the highest classification score with the assessment of the human paleographer (PAS). The obtained accuracy is 36%.

Next, building on our intuition that an unmarked descriptor might actually be present in a given handwriting, we employ a matrix-based method often used for imputing missing data. The method approximates the observation matrix (in our case, the size equals the number of descriptors times the number of documents) as a low-rank matrix using Singular Value Decomposition (SVD). The missing elements are taken directly from the corresponding elements of the approximation (Hastie et al., 1999). Retraining and employing SVM to the obtained descriptor vectors yields only a slight improvement in performance to 37%.

The imputed values are real-valued. We aim to choose an appropriate cut-off threshold for each of the categories. To this end, for every class, we rank all documents by the classification score obtained by the specific SVM model. Then, for each descriptor, we ask what would be the threshold least likely to occur by chance (see "Technical details" below).

Applying all the per-descriptor thresholds, a new set of binary exists/does-not-exist representations is obtained for each handwriting, and SVM-based

classification is applied as before. This new method shows a remarkable increase in performance, and 49% of the documents are classified correctly.

To further illustrate the effectivness of the method, we consider not just the first classification provided by the system, but the top three. SVM on the original descriptor vectors provides the correct answer as one of the top-three classes 78% of the time. Using the SVD based imputation method, the performance remains 78%. Finally, using the new method, the performance improves to 84%.

Technical details

The underlying method compares two ranked lists and returns the pair of thresholds which are the least likely to occur by chance. In our case, one list is a list of classification scores for a specific category, and the other contains imputed scores for a given descriptor. Both ranked lists are of the same length – n – which is the number of handwritings.

Let x and y be two vectors in $\mathbb{R}^n$. Applying a threshold to either vector divides the elements of this vector into two groups. A natural association between x and y would capture whether there exist thresholds such that the sets of obtained indices significantly overlap.

The hypergeometric distribution $f(k;n,i,j)$ captures the probability of obtaining a certain intersection size k between two sets X and Y of given sizes $i := |X|$, and $j := |Y|$, where the elements of the two sets are drawn randomly from the set $1 \ldots n$: $P(|X \cap Y| = k) = f(k;n,i,j)$.

To evaluate the statistical significance of a certain intersection size, we consider the probability of obtaining an intersection at least as large by random drawing from $1 \ldots n$ two sets of sizes i and j. To that end we employ the hypergeometric cumulative distribution function $F(k;n,i,j)$, which measures the probability of obtaining an intersection size of up to k: $F(k;n,i,j) = \sum_{c=0}^{k} f(c;n,i,j)$. The statistical significance we consider (probability of an intersection size of at least k) is therefore given by the tail probability: $G(k;n,i,j) = 1 - F(k-1;n,i,j)$.

Given a vector $x \in \mathbb{R}^n$ of unique values, there are $n + 1$ possible threshold-based subsets of the indices $1 \ldots n$, i.e., sets X such that for every $p \in X$, $x_q < x_p$ implies $q \in X$. Each such subset is uniquely identified by its size. Denote these subsets by $X_0, X_1, \ldots, X_n$ such that $|X_i| = i$.

Considering also the vector y, ordered in a similar manner and giving rise to the ordered subsets of indices $Y_0, \ldots, Y_n$. Let $I \in \mathbb{R}^{n \times n}$ be the matrix such that $I_{i,j} = |X^i \cap Y^j|$.

We define the matrix P where $P_{i,j}$ is the probability of obtaining an intersection size of at least $I_{i,j}$ for sets of sizes i and j, when randomly drawing indices from $1 \ldots n$: $P_{i,j} = G(I_{i,j},n,i,j)$.

We seek thresholds whose values produce the minimal value of P, i.e., they produce the subsets of sizes i and j for which the following minimum is obtained: $min_{i,j} P_{i,j}$.

For n documents, a naive computation of the matrix I requires $O(n^3)$. This can be improved to $O(n^2)$ by considering the lists of indices obtained by sorting x and y.

Let C be the matrix defined such that $C_{i,j} = 1$ if the jth sorted index of y is in the first i sorted indices of x. C can be computed from $x, y \in \mathbb{R}^n$ in time and

storage complexity of $O(n^2)$. The following lemma shows that I can be computed from C in a similar time complexity by performing cumulative sum over the rows of C.

Lemma 1. *For every $x, y \in \mathbb{R}^n$, and for C and I as above, $I_{i,j} = \sum_{k=1}^{j} C_{i,k}$.*

Once I is computed, P is readily evaluated based on the hypergeometric cumulative distribution function. An efficient algorithm is given in (Berkopec, 2007), which has as many iterations as $min(n - i, n - j)$. Using the identity $F(k; n, i, j) = 1 - F(n - k - 1; n, m - i, j)$ (Riordan, 1968), the number of iterations can be further reduced to $min(n - i, n - j, i, j)$. Still, considering that $P_{i,j}$ is evaluated for all $i = 1 \ldots n$ and $j = 1 \ldots n$ this is computationally demanding for large n.

The following lemma can be used to reduce the number of evaluations of the hypergeometric cumulative distribution function. It states that by examining the elements of the matrix C around the location i, j, we are able to determine whether $P_{i,j}$ can potentially obtain the minimal value out of all elements of P.

Lemma 2. *Given any vectors x and y, let C, I, and P be defined as above, then if $P_{i,j}$ is a minimal value of the matrix P the following two conditions hold: (a) $C_{i,j} = 1$; and (b) $j < n \Rightarrow C_{i,j+1} = 0$.*

Experimentally it is found that using lemma 2, between 75% and 85% of the entries of the matrix P need not be computed, where the larger n is, the higher the ratio of discarded entries.

Discussion

Descriptor-based approaches are a key component in shifting paleography from an authoritative discipline to an evidence-based one in which expert rulings can be explained. In an evidence-based approach, decisions should be based on descriptors in the manuscript which can be readily verified by other experts. It should be noted that the ability to rely on concrete evidence does not mean that classification accuracy is improved. The classification of the authoritative expert who is free from the need to explain herself would probably be at least as accurate, if not very much more. Thus, in order to achieve high levels of performance, it is crucial to have accurate decision rules and models on top of the descriptors.

It is also worth observing that none of these systems are truly objective. The premise of the approach taken here is that different people will inevitably make different decisions when selecting and recording descriptors: that the input data in any system is necessarily the result of selection and human decisions with everything that this entails. Indeed, the method outlined in this paper relies on an initial set of descriptors which have themselves been selected by experts, and so any bias in that original selection will necessarily be reflected in the descriptors which it predicts. Nevertheless, it does help to reduce the degree of variation when different people are entering data into a system, as normally happens in large projects in the Digital Humanities. As well as improving classification, it can also suggest descriptors that have been overlooked, and so project members who

are entering the data can then go and check their work. In this respect the method applies much more widely than simply to paleography, since the problem of consistency in selection across a team is widespread.

Building on the observation that unmarked descriptors are occasionally missing for the "wrong" reasons, we are able to improve classification accuracy significantly. The method relies on several underlying assumptions that should be considered. First, by means of the low-rank approximation, the prediction of the missing descriptors is based on past correlations between the various descriptors. Therefore, a unique configuration of descriptors would be augmented to become a more conventional one, possibly losing valuable information. Second, by means of examining the correlations between descriptors and class memberships, our method assumes that the descriptors are discriminative. As a future direction we can apply our method more selectively, only to descriptors that appear (on the training data) to be informative.

Fig. 1: (left) Hand G640-3 [http://image.ox.ac.uk/images/bodleian/msjunius11/226.jpg]. The system predicted r long foot and y hooked tail which were indeed missing, and e tall-lig which would fit the style but is missing. (right) Hand G314-1 [http://digipal.cch.kcl.ac.uk/hand/page/216/]. y round undotted predicted by the system is unlikely. eth turned-down tip is somewhat more likely; ash convex-left was indeed missing from the annotation.