Trimming Prototypes of Handwritten Digit Images with Subset Infinite Relational Model

Tomonari Masada[1] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, 852-8521 Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
takasu@nii.ac.jp

Abstract. We propose a new probabilistic model for constructing efficient prototypes of handwritten digit images. We assume that all digit images are of the same size and obtain one color histogram for each pixel by counting the number of occurrences of each color over multiple images. For example, when we conduct the counting over the images of digit "5", we obtain a set of histograms as a *prototype* of digit "5". After normalizing each histogram to a probability distribution, we can classify an unknown digit image by multiplying probabilities of the colors appearing at each pixel of the unknown image. We regard this method as the baseline and compare it with a method using our probabilistic model called Multinomialized Subset Infinite Relational Model (MSIRM), which gives a prototype, where color histograms are clustered column- and row-wise. The number of clusters is adjusted flexibly with Chinese restaurant process. Further, MSIRM can detect *irrelevant* columns and rows. An experiment, comparing our method with the baseline and also with a method using Dirichlet process mixture, revealed that MSIRM could neatly detect irrelevant columns and rows at peripheral part of digit images. That is, MSIRM could "trim" irrelevant part. By utilizing this trimming, we could speed up classification of unknown images.

Keywords: Bayesian nonparametrics, prototype, classification

1 Introduction

This paper considers image classification. While there are a vast variety of methods, we focus on *prototype*-based methods. We construct a prototype for each image cate-

gory and classify an unknown image to the category whose prototype is the most similar. In this paper, we describe prototypes with probability distributions and classify an unknown image to the category whose prototype gives the largest probability.

We assume that all images are of the same size, say N_1 by N_2 pixels, because we consider *handwritten digit images* in this paper. We can obtain a set of color histograms by counting the number of occurrences of each color at each pixel. Consequently, we obtain $N_1 N_2$ color histograms, each at a different pixel. When we conduct this counting over the images of the same category, e.g. the images of handwritten digit "5", the resulting set of histograms gives a color configuration specific to the category and can be regarded as a *prototype* of the category. Formally, we describe each prototype with parameters $\{g_{ijw}^h\}$, where g_{ijw}^h is a probability that the wth color appears at the 2D pixel location (i,j) for $w = 1,\dots,W$, $i = 1,\dots,N_1$, $j = 1,\dots,N_2$, where W is the number of different colors. The superscript h is a category index. By using the parameters, we can calculate the log probability of an unknown image as $\sum_{i,j,w} n_{ijw} \ln g_{ijw}^h$, where n_{ijw} is 1 if the image has the wth color at the pixel (i,j) and 0 otherwise. Based on the obtained probabilities, a category for the image can be determined by $\arg\max_h \sum_{i,j,w} n_{ijw} \ln g_{ijw}^h$. We regard this method as the baseline method and would like to improve it in terms of *efficiency*.

Any prototype the baseline gives has $N_1 N_2 W$ parameters, whose number can be reduced by *clustering* histograms. We propose a new probabilistic model called Multinomialized Subset Infinite Relational Model (MSIRM) for clustering. MSIRM clusters histograms column- and row-wise. Denote the numbers of column and row clusters as K_1 and K_2, respectively. In MSIRM, each pixel is assigned to a pair of column and row clusters, and the colors of the pixels assigned to the same pair of column and row clusters are assumed to be drawn from the same distribution. Consequently, we can reduce the complexity of prototypes from $N_1 N_2 W$ to $K_1 K_2 W$. MSIRM determines K_1 and K_2 flexibly with Chinese restaurant process (CRP) [4]. MSIRM has an important feature: it detects columns and rows *irrelevant* for constructing prototypes. This feature can speed up classification, because we can reduce execution time of classification by skipping irrelevant columns and rows. The skipping techni-

cally means giving probability one to all pixels in irrelevant columns and rows. We will show that the skipping lead to only a small degradation in classification accuracy.

The rest of the paper is organized as follows. Section 2 gives preceding proposals important for us. Section 3 provides details of MSIRM. Section 4 presents the results of our comparison experiment. Section 5 concludes the paper with discussions.

2 IRM and SIRM

We propose MSIRM as an extension of Subset Infinite Relational Model (SIRM) [1], which is, in turn, an improvement of Infinite Relational Model (IRM) [2].

Assume that we have N_1 entities of type T_A and N_2 entities of type T_B, and that a binary relation R is defined over a domain $T_A \times T_B$. The relation can be represented by an N_1 by N_2 binary matrix as the left most panel of Fig. 1 shows. IRM discovers a column- and row-wise clustering of entities so that the matrix gives a relatively clean block structure when sorted according to the clustering as the center panel of Fig. 1 shows. IRM can determine the numbers of column and row clusters flexibly with CPR. Further, IRM associates each block, enclosed by thick lines in the center panel of Fig. 1, with a binomial distribution determining the probability that the pairs of entities in the block fall under the relation R. While IRM can be extended to the cases where there are more than two types of entity, we do not consider such cases.

When constructing a prototype from the images of the same category, we can consider a relation between pixel columns and rows by taking an IRM-like approach, because contiguous pixels are likely to give similar color distributions. However, IRM is vulnerable to noisy data. Therefore, a mechanism for detecting *irrelevant* columns and rows is introduced by SIRM, where clustering is conducted only on a *subset* of columns and rows, i.e., on columns and rows other than irrelevant ones. In SIRM, we flip a coin at each column and row to determine whether relevant or not. Irrelevant pixels are bundled into the same group, as the right most panel of Fig. 1 depicts with red cells. However, both IRM and SIRM can only handle binary data. Therefore, we "multinomialize" SIRM for handling multiple colors and obtain our model, MSIRM.

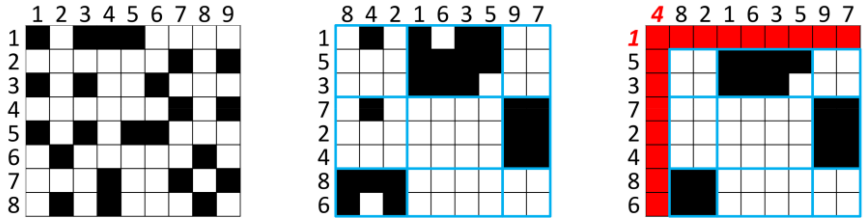

Fig. 1. Clustering of a binary relation (left) by IRM (center) and by SIRM (right). Each cluster is enclosed by thick lines. Red cells in the right most panel correspond to irrelevant pixels.

3 MSIRM

We below describe how observed data are generated by MSIRM.

1. Draw the parameter λ_1 of a binomial $\mathrm{Bi}(\lambda_1)$ from a Beta prior $\mathrm{Be}(a_1, b_1)$. Then, for each column, draw a 0/1 value from $\mathrm{Bi}(\lambda_1)$. Let r_{1i} denote the value for the ith column, which is irrelevant if $r_{1i} = 0$ and is relevant otherwise.
2. Draw the parameter λ_2 of a binomial $\mathrm{Bi}(\lambda_2)$ from a Beta prior $\mathrm{Be}(a_2, b_2)$. Then, for each row, draw a 0/1 value from $\mathrm{Bi}(\lambda_2)$. Let r_{2j} denote the value for the jth row, which is irrelevant if $r_{2i} = 0$ and is relevant otherwise.
3. Draw the parameters $\phi_{kl1}, \ldots, \phi_{klW}$ of a multinomial distribution $\mathrm{Mul}(\boldsymbol{\phi}_{kl})$ from a Dirichlet prior $\mathrm{Dir}(\beta)$. $\mathrm{Mul}(\boldsymbol{\phi}_{kl})$ is the multinomial for generating colors of the pixels belonging to the kth column and, at the same time, to the lth row clusters.
4. Draw the parameters $\psi_1, \ldots, \psi_W$ of a multinomial $\mathrm{Mul}(\boldsymbol{\psi})$ from a Dirichlet prior $\mathrm{Dir}(\gamma)$. $\mathrm{Mul}(\boldsymbol{\psi})$ is the multinomial for generating colors of the irrelevant pixels.
5. For each relevant column, draw a cluster ID based on a Chinese restaurant process $\mathrm{CRP}(\alpha_1)$. We introduce a latent variable z_{1i}, which is equal to k if the ith column is relevant and belongs to the kth column cluster.
6. For each relevant row, draw a cluster ID based on a Chinese restaurant process $\mathrm{CRP}(\alpha_2)$. We introduce a latent variable z_{2j}, which is equal to l if the jth row is relevant and belongs to the lth column cluster.
7. For each pixel (i, j), draw a color from the multinomial $\mathrm{Mul}(\boldsymbol{\psi})$ if $r_{1i} r_{2j} = 0$ (i.e., the pixel is irrelevant) and from the multinomial $\mathrm{Mul}\left(\boldsymbol{\phi}_{z_{1i} z_{2j}}\right)$ otherwise.

We adopt Gibbs sampling technique for inferring posterior distribution of MSIRM. For each column, we update the relevant/irrelevant coin flip r_{1i} and the cluster assignment z_{1i} based on the following posterior probability:

$$p(z_{1i}, r_{1i} | \boldsymbol{X}, \boldsymbol{Z}^{\backslash 1i}, \boldsymbol{R}^{\backslash 1i}) \propto p(\boldsymbol{X}^{+1i} | z_{1i}, r_{1i}, \boldsymbol{X}^{\backslash 1i}, \boldsymbol{Z}^{\backslash 1i}, \boldsymbol{R}^{\backslash 1i}) p(z_{1i}, r_{1i} | \boldsymbol{Z}^{\backslash 1i}, \boldsymbol{R}^{\backslash 1i}),$$

where the details of the two terms on the right hand side are omitted due to space limitation. The derivation is almost the same with that for SIRM [1]. We also conduct a similar update for each row. Additionally, we update the hyperparameters of the Dirichlet and the Beta priors by Minka's method[1].

4 Experiment

Our experiment compared MSIRM with the baseline and also with a method using Dirichlet process mixture of multinomial (DP-multinomial) [3] for clustering histograms. Our target was MNIST data set[2], consisting of 60,000 training and 10,000 test images. All images are 28 by 28 pixels in size, i.e., $N_1 = N_2 = 28$. We quantized 8 bit gray-scaled colors into 4 bit colors uniformly, i.e., $W = 16$.

Let n_{ijw}^h be the number of times the wth color appears at the pixel (i, j) in the training images of category h, where h ranges over ten digit categories. The baseline method calculates a probability g_{ijw}^h that the wth color appears at the pixel (i, j) in the images of category h as $g_{ijw}^h = (n_{ijw}^h + \eta_w) / \sum_w (n_{ijw}^h + \eta_w)$, where a Dirichlet smoothing with the parameters $\eta_1, \ldots, \eta_W$ is applied. We determine the category of an unknown image by $\arg\max_h \sum_{i,j} \hat{n}_{ijw} \ln g_{ijw}^h$, where $\hat{n}_{ijw}$ is 1 if the wth color appears at the pixel (i, j) in the unknown image and is 0 otherwise. For the DP-multinomial method, we set g_{ijw}^h to a posterior probability estimated by a Gibbs sampling [3] and determine the category of an unknown image in the same manner with the baseline method. For MSIRM, we determine the category of an unknown image as follows:

---

[1] http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/

[2] http://yann.lecun.com/exdb/mnist/

$$\arg\max_{h} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \hat{n}_{ijw} \left\{ r_{1i} r_{2j} \ln \frac{n^h_{z_{1i} z_{2j} w} + \beta^h_w}{\sum_w (n^h_{z_{1i} z_{2j} w} + \beta^h_w)} + (1 - r_{1i} r_{2j}) \frac{q^h_w + \gamma^h_w}{\sum_w (q^h_w + \gamma^h_w)} \right\},$$

where n^h_{klw} is the number of times the wth color appears at the pixels belonging to the kth column and the lth row clusters in the training images of category h. q^h_w is the number of times the wth color appears at the irrelevant pixels of the training images of category h. β^h_w and γ^h_w are the hyperparameters for category h.

We give results of the experiment. Let K be the number of clusters given by DP-multinomial. With respect to the complexity of prototypes, DP-multinomial was superior to MSIRM. While $K_1 \approx 20$ and $K_2 \approx 20$ for MSIRM, $K \approx 85$ for DP-multinomial with respect to all digits. That is, $K < K_1 K_2$ for all digits.

Classification accuracies of the baseline, DP-multinomial, and MSIRM were 0.840, 0.839, and 0.837, respectively. While the accuracies were far from the best reported at the Web site of the data set, they were good enough for a meaningful comparison. MSIRM and DP-multinomial gave almost the same accuracies with that of the baseline. Therefore, it can be said that we could reduce the complexity of prototypes by clustering histograms without harming clustering accuracy.

While DP-multinomial and MSIRM showed no significant difference in terms of accuracy, MSIRM could neatly detect irrelevant columns and rows at peripheral part of images. Fig. 2 visualizes the prototypes obtained by MSIRM. We mixed grayscale colors linearly by multiplying their probabilities at each pixel and obtained the visualization in Fig. 2. The red-colored area corresponds to irrelevant columns and rows. Astonishingly, MSIRM precisely detected the area irrelevant for digit identification.

By utilizing this "trimming" effect, we could speed up classification. We skipped irrelevant columns and rows when we calculated the log probabilities of unknown images. Technically, this means that we assigned probability one to all irrelevant pixels. For the prototypes in Fig. 2, we could skip 32.2% of the $28^2 \times 10 = 7{,}840$ pixels. This led to 32.2% speeding up of classification. The achieved accuracy was 0.819, which shows a small degradation from 0.837. Therefore, it can be said that MSIRM could speed up classification with only a small degradation in accuracy.

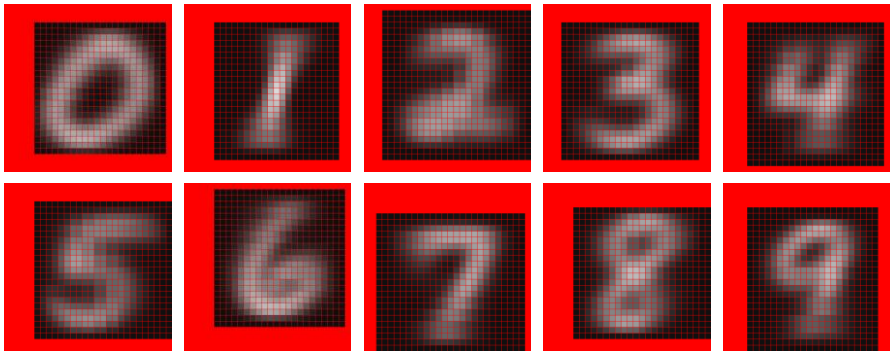

Fig. 2. Irrelevant portion (red-colored pixels) trimmed out by MSIRM from each prototype.

5 Conclusion

This paper proposes a prototype-based image classification using MSIRM. We could neatly detect irrelevant pixels and could speed up classification by skipping the irrelevant pixels. This led to only a small degradation in accuracy. While DP-multinomial was comparable with MSIRM in accuracy and was superior to MSIRM in reduction of the complexity of prototypes, it has no mechanism for detecting irrelevant pixels.

We have a future plan to extend MSIRM for realizing a clustering of training images by introducing an additional axis aside from the column and the row axes.

References

1. Ishiguro, K., Ueda, N., Sawada, H.: Subset infinite relational models. JMLR W&CP 22 (AISTATS 2012), 547-555 (2012)
2. Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. in Proc. of AAAI'06, pp. 381-388 (2006)
3. Neal, R. M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. J. Comput. Graph. Statist., 9(2), 249-265 (2000)
4. Pitman, J.: Combinatorial stochastic processes. Notes for Saint Flour Summer School (2002)