Sample Clustering for Fast Classification by Using the Mean Shift Procedure

Liang Lie-quan
Guangdong Provincial Key Lab of E-Commerce Marketing Application Technology
Guangdong Commerce College
Guangzhou, P. R China, 510320le
Guangzhou 510320 P. R China
lianglq@gdcc.edu.cn

Liang Ying-hong
Guangdong Provincial Key Lab of E-Commerce Marketing Application Technology
Guangdong Commerce College
Guangzhou, P. R China, 510320le
Guangzhou 510320 P. R China
lyh@gdcc.edu.cn

***Abstract*—Most classification methods are limited by speed particularly when the training data set is large, such as artificial neural networks (ANNs) and support vector machines (SVMs). In this article, we explore the possibility of utilizing the Mean Shift algorithm, which is a mode seeking procedure that estimates the gradient of the data density, to decrease the sample size. We found that in a large number of samples to be trained, most samples can be clustered into a small number of mode centroids (extreme values of density), therefore, the original samples can be reduced by means of using the results of the Mean Shift procedure. To verify the validity of this method, several classifiers including the linear discriminant analysis (LDA), k nearest neighbor (kNN) and SVMs have been tested. Experimental results prove that when the parameters are selected appropriately, the proposed method is capable of reducing the computational complexity of above classification methods, with minimum effects on the classification accuracy.**

Keywords- sample reductio; mean shift; classification methods; mode seeking; sample selection

I. INTRODUCTION

In recent years, many classification methods have been introduced in the field of statistical pattern recognition (STP). The main task of STP is to use the pre-labeled training samples to establish decision boundaries in the feature space which separate patterns belonging to different classes [6-7]. Thus, the performance of classification methods is highly dependant on the patterns in the training samples. Although most methods have shown promising performance in classification, they are limited by speed particularly when the training data set is large. Another possible problem caused by a large sample set is over-fitting (the classifier which correctly classifies the training samples may not correctly label an independent set of testing samples). As a matter of fact, in a larger sample set, not all samples are relevant to build the proper decision function. Hence removing any irrelevant or unimportant samples in the training set may have no effect on the classification accuracy. Such a technique is called sample selection or sample reduction [8-10].

In [8], a sample reduction method using clustering techniques to remove non-relevant samples in deciding the decision boundary for SVM is proposed. The performance is excellent when applied to SVM classifiers. In [9], the authors explore the possibility of using a reduced number of samples as centroids to construct efficient SVM-like machines. In [10], the authors show the impact of sample reduction on the process of feature extraction for supervised learning. However, all those methods mentioned above are not designed for general proposal.

In this paper, a Mean Shift based sample clustering approach for fast classification is proposed. The basic idea of this method is to cluster the original samples into a small number of mode centroids and use these centroids as new samples. We use several techniques including the LDA, kNN and SVMs to verify the validity of this method. We also examine how the two parameters (kernel bandwidth and admissible error) in the Mean Shift algorithm affect the number of samples removed, as well as the correctly classified rates.

The remainder of this paper is organized as follows. Section 2 introduces our method for sample reduction in detail. Section 3 demonstrates the impact of parameters on sample reducing. Experimental results are given in Section 4. Finally, the conclusions are set out in Section 5.

II. THE APPROACH FOR REDUCING THE NUMBER OF TRAINING SAMPLES

A. The Mean Shift clustering procedure

First, The Mean Shift algorithm which originated from the density gradient estimation [1], did not receive attention until the paper [2] proposed that it was a mode-seeking procedure, which detects the mode of a data set using an iterative procedure. The author in [2] also extended the Mean Shift algorithm with kernel functions. In [3, 4], the authors gave detailed discussion on the relationship between the Mean Shift procedure and the non-parametric kernel estimation (non-parametric multivariate kernel density regression).

Let $\{x_i\}_{i=1}^{N}$ be a set of samples in the feature space, where x is a d-dimensional vector. The mean shift without kernel function at x is:

978-0-7695-3643-9/09 $25.00 © 2009 IEEE
DOI 10.1109/ISECS.2009.72

$$M(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - x) \tag{1}$$

$M(x)$ is in the gradient direction at x of the density estimation. Obviously, here the data points have the same contributions for the calculation regardless of their distance from x. To overcome this shortcoming, the kernel function is introduced into the mean shift computation. The mean shift with kernel function can be expressed as:

$$M(x) = \frac{\sum_{i=1}^{N} K_H(x_i - x)w(x_i)(x_i - x)}{\sum_{i=1}^{N} K_H(x_i - x)w(x_i)} \tag{2}$$

where $K_H = |H|^{-1/2} K(H^{-1/2}x)$ is the kernel function, used to influence the distributions of data points according to their distance from x, $w(x)$ are the normalized weights where $\sum_{i=1}^{N} w(x) = 1$, H is a $d \times d$ bandwidth matrix to control the smoothness of the density estimation. To avoid a costly matrix inversion, H is replaced by an identity matrix $H = h^2 I$

$$M(x) = \frac{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)(x_i - x)}{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)} \tag{3}$$

Furthermore, Eq. (3) can be transformed as:

$$M(x) = \frac{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)x_i}{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)} - x \Rightarrow m(x) = \frac{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)x_i}{\sum_{i=1}^{N} K(\frac{x_i - x}{h})w(x_i)} \tag{4}$$

$m(x)$ is the sample mean with kernel function. Assuming all the data points have the same weight, $m(x)$ can be simplified as:

$$m(x) = \frac{\sum_{i=1}^{N} K(\frac{x_i - x}{h})x_i}{\sum_{i=1}^{N} K(\frac{x_i - x}{h})} \tag{5}$$

In addition, using the Gaussian kernel function $K(x) = (\sqrt{2\pi}h)^{-1} e^{-1/2(x/h)^2}$, Eq. (5) becomes:

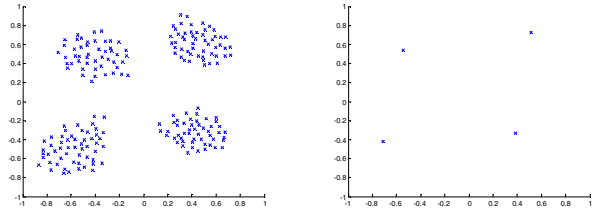
$$m(x) = \frac{\sum_{i=1}^{N} e^{-\frac{(x_i - x)(x_i - x)'}{2h^2}} x_i}{\sum_{i=1}^{N} e^{-\frac{(x_i - x)(x_i - x)'}{2h^2}}} \tag{6}$$

Let $P \subset X$ be a set of mode centroids, from each x in X, the m(x) can be computed with iterations, which make $P \leftarrow m(P)$. This form of iterations is called the Mean Shift algorithm (procedure) [2]. It is easy to understand that the Mean Shift procedure is indeed an estimation process of the gradient of the data density.

Thus, due to the repeated movement of m(x) to the peak value, the Mean Shift based clustering method is an iterative process. To make the iterative procedure converge rapidly, a minimum threshold ε for the allowed error deviation in consecutive iterations is given, which is called an admissible error, the calculation of m(x) stops when || m(x) - x || <= ε. The convergence proof of the Mean Shift algorithm is given in [3]. The authors also give the sufficient conditions for convergence.

Now we consider the sample reduction problem. As discussed above, when the number of training samples is huge, it will be computationally expensive to use all of them for training. Actually, we found that most samples can be clustered into a small number of mode centroids. Therefore, we can remove the high-density clusters of training samples in the feature space, and use their mode centroids to generate a new sample set.

Fig.1 gives a simple example of sample clustering by using the Mean Shift procedure. The original sample set is composed of 208 two dimensional points which can be clustered into 4 high-density clusters. Using the Mean Shift procedure with a Gaussian Kernel function (see Eq. (6), h = 0.1, ε = 0.4) the distribution of mode centroids shown in Fig. 1(b) is obtained, which clearly identifies those 4 high-density clusters in the original sample set are represented by their mode centroids.



(a) Distribution of the original sample (b) Distribution of the mode centroids

Figure 1. An example of sample clustering via the Mean Shift procedure

Since the reduced sample set is not a subset, but the mode centroids of the original training set, the proposed sample reduction approach is in fact a sample transformation method. Although the training sample size can be remarkably decreased by this method, we face the problem of deciding how many new samples are needed to maintain the boundaries of different classes, so that the sample reducing operation will not seriously affect the classification results. The parameters of this algorithm are how to affect the number of reduced samples will be discussed in Section 3.

B. Algorithm description

We assume the representation of a training sample set S

$$S = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \subseteq (X \times Y)^n \tag{7}$$

where each sample item $s_i = (x_i, y_i)$ consists of a d-dimensional feature vector $x_i \in X \subseteq R^d$ and a corresponding class label $y_i \in Y = \{1, ..., c\}$. Each class set Si ($1 \le i \le c$) has ni elements

and the total number of samples is $n = \sum_{i=1}^{c} n_i$. Similarly, when the sample clustering is finished, the new sample set can be expressed as

$$S' = ((x_1', y_1'), (x_2', y_2'), ..., (x_l', y_l')) \subseteq (X \times Y)^l \quad (8)$$

where each class set S_i' has li elements and the total number of new samples is l (l < n).

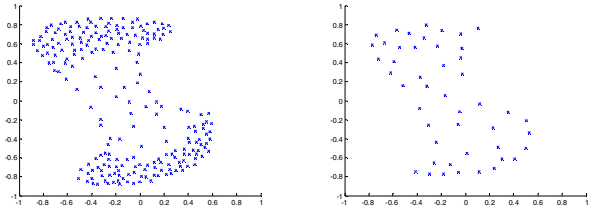
So far, the proposed sample reducing method can be summarized as follows. Given a large training set S, for each class set Si, we use the Mean Shift procedure to seek its mode centroid set S_i' . Finally, all the centroid sets are composed of a new sample set S' . The proposed algorithm works as follows:

Input: S, the original sample set; S_i ($1 \le i \le c$), the subset of class i; n, the total number of the original samples; n_i, the number of class i.

Output: $S_i' = \{\}$, the new sample set; $l = 0$, the total number of the new samples; $l_i = 0, S_i' = \{\}$ ($1 \le i \le c$).

(1) Set the admissible ε and the bandwidth h.
(2) For each class set S_i in S
(3) For each element x_j in S_i
(4) Calculate $m(x_j)$ using Eq. (6);
(5) If $\| m(x_j) - x_j \| > \varepsilon$, let $x_j = m(x_j)$, repeat Step (4);
(6) For each element x_k' in S_i'
(7) If $\| m(x_j) - x_k' \| > \varepsilon$, then $S_i' = S_i' \cup \{m(x_j)\}$, $l_i = l_i + 1$;
(8) End for
(7) End for
(8) End for
(9) $S' = S_1' \cup S_2' \cup ... \cup S_c'$, $l = l_1 + l_2 + ... + l_c$.

Fig. 2 displays an example of sample reduction by using a small number of mode centroids as new samples, by means of our sample clustering algorithm (Gaussian kernel, h = 0.1, ε = 0.1). The total number of the original samples is 223, while the number of the new samples is 47. As shown in Fig. 2(b), the original samples have been reduced by 79%. However the distribution boundary of the original set was approximately maintained. According to some classification methods, such as SVMs, samples close to the decision boundaries have a higher likelihood of being a support vector. Therefore it is very important to keep the samples close to the borders, in order to maintain the same accuracy.



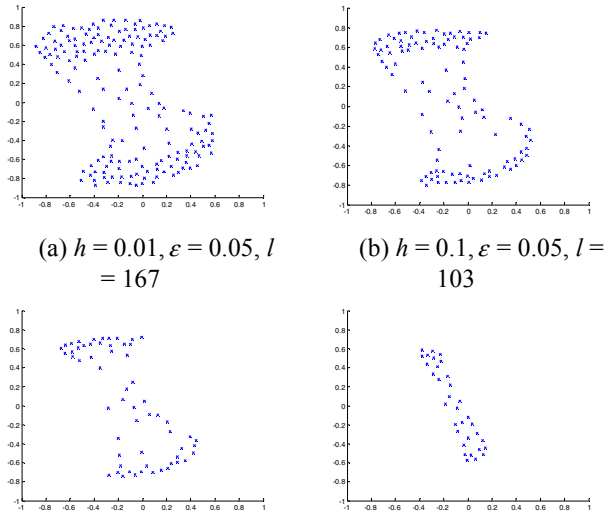
(a) Original sample set (b) New sample set

Figure 2. Sample reduction by using a small number of mode centroids as new samples

III. THE IMPACT OF PARAMETERS ON SAMPLE DISTRIBUTION

A. Bandwidth

The Mean Shift procedure comes from the Parzen window technique [5] that belongs to the non-parametric statistics. Therefore, similar to the non-parametric density estimation, both the sample size and kernel bandwidth can influence the estimation precision. In general, given a sample size, using a small bandwidth (narrow kernel) yields a ragged density, thus a great deal of mode centroids are obtained, while using a big bandwidth (wider kernel) results in a smoother density and a smaller number of mode centroids. A sample clustering example for different bandwidth values is shown in Fig. 3. The original training set (see Fig. 2(a)) was processed with the Mean Shift procedure employing Gaussian kernels using various bandwidth values (0.01, 0.1, 0.2, and 0.5). It is obvious from the figures that the number of mode centroids reduced as the bandwidth increased. We can also notice that using a wider kernel leads to a significantly narrowed distribution (see Fig. 3(d)). As we pointed out above, keeping the distribution borders is very important to maintain the classification accuracy. So, obviously we cannot control the number of reduced samples through modifying the bandwidth value.



(a) $h = 0.01, \varepsilon = 0.05, l = 167$ (b) $h = 0.1, \varepsilon = 0.05, l = 103$

(c) $h = 0.2, \varepsilon = 0.05, l = 59$ (d) $h = 0.5, \varepsilon = 0.05, l = 33$

Figure 3. An example of sample clustering for different bandwidths

To overcome the problem that when the bandwidth is not selected appropriately, the distribution of mode centroids will vary much and therefore can't be used for training, we use a fixed bandwidth in our algorithm. Referring to the method of bandwidth selection of the non-parametric density estimation discussed in [11], a simplest way to choose h with a Gaussian kernel can be performed as:

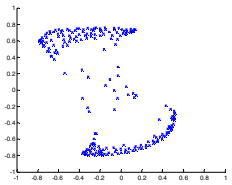
$$h_j^{opt} = \left(\frac{4}{2d+1}\right)^{1/(d+4)} \hat{\sigma}_j n^{-1/(d+4)} \quad (9)$$

where d is the number of dimensions, n is the number of samples, $\hat{\sigma}_j^2$ is covariance for the jth distribution.

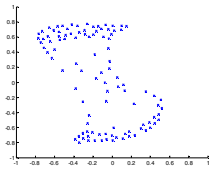
Thus, in our algorithm, we first calculate the optimal bandwidth using Eq. (9) for each class, and then perform the Mean Shift clustering procedure to find their mode centroids.

B. *Admissible error*

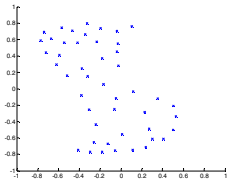
Another parameter that can affect the sample reduction result is the admissible error ε . In our algorithm, the admissible error has two effects: First, it determines the convergence speed. Second, it controls the minimum distance between two mode centroids. We can see how the admissible error affects the sample clustering in Fig. 4. Obviously, the number of the new sample set trends to decrease as the admissible increases. However, the pattern of the distribution is maintained. In Fig. 4(d), even though only 11 mode centroids were found, we can still obtain the similar distribution pattern with the original one (see Fig. 2(a)). Thus, compared with the bandwidth, the admissible error is more suitable for controlling the number of reduced samples.



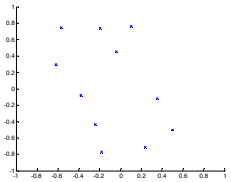
(a) $h = 0.1, \varepsilon = 0.01, l =$ 214



(b) $h = 0.1, \varepsilon = 0.05, l =$ 103



(c) $h = 0.1, \varepsilon = 0.1, l =$ 47



(d) $h = 0.1, \varepsilon = 0.3, l =$ 11

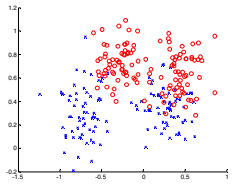
Figure 4. An example of sample clustering for different admissible errors

IV. EXPERIMENTS AND DISCUSSION

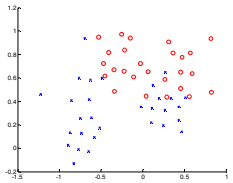
A. *Experiment on a two-dimensional dataset*

For the purpose of illustration of the algorithm, as shown in Figs. 5(a) and 5(c), we use a database of two classes called Riply from the STPRtool [12]. In the database, the number of the training set and the number of the test set are 250 and 1000, respectively. All the samples are two-dimensional points. Fig. 5(b) shows the sample reducing result using the proposed algorithm. The total number of the new training set is 56.

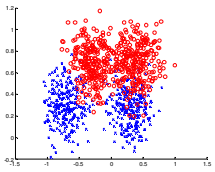
The experiment is performed by using the LDA, kNN(k=8) and SVM classifiers. Note that here we also use the codes in STPRtool available at the website [12]. The trained classifiers are visualized in Fig. 5. Fig. 5(a) shows the linear classifiers trained by the LDA algorithm, which are visualized as separating hyperplanes. Fig. 5(b) shows the decision boundaries obtained from the kNN classifiers. Fig. 5(c) displays the decision boundaries of the SVM classifiers. From these figures, it is easy to see that the training results form the reduced set is extremely close to the results form the whole training set.



(a) The original training sample set, $n = 250$

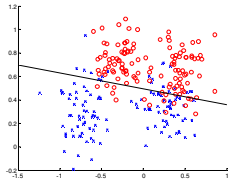


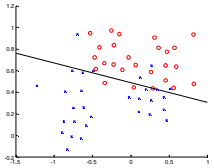
(b) The reduced training sample set, $l = 56$



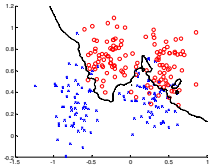
(c) The test set, $n =$ 1000

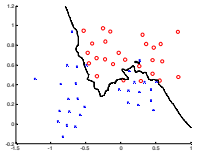
Figure 5. An example dataset for a nonlinearly separable two-class problem



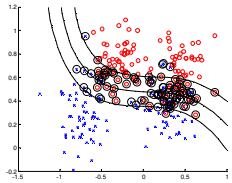


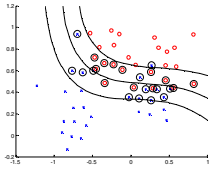
(a) Linear classifiers trained by the LDA algorithm.





(b) Decision boundaries of the *k*NN classifiers ($k = 8$)





(c) Decision boundaries of the SVM classifiers (Gaussian kernel, bandwidth is 1)

Figure 6. Figures showing the different classifiers trained on the Riply database and their reduced sample sets (The training results of the original sample set and reduced sample set are shown in the left and right figures, respectively.)

In addition, we compare the reduced set with the whole set in the terms of classification accuracy. Results can be seen in Table 1. In all three cases, there was no significant change in the classification rate. For LDA and kNN, the classification rates have improved by 0.2 and 2.3, respectively. The

improvement possibly results from the over-fitting of the original training samples. For SVM, the slight decrease of the classification rate is because of the reduction of the support vectors.

TABLE I. COMPARISON OF CLASSIFICATION ACCURACIES

	Original set			Reduced set		
	LDA	*k*NN	SVM	LDA	*k*NN	SVM
Correctly classified (%)	89.2	88.4	90.6	89.4	90.7	89.6

B. Experiment on the Letter Image Recognition Dataset

In this experiment, we use the Letter Image Recognition Dataset [13]. This dataset contains 20,000 16-dimensional samples of 26 classes. 16,000 samples are used as the training set and the remaining 4,000 are the test set. The new training set obtained by our sample reduction algorithm consists of 7183 samples. Two classifiers are applied to the training sets, including the kNN (k = 4) and SVM (Gussian kernel function, the bandwidth is 1). This experiment has been repeated 5 times in order to get average results. Tables 2 and 3 show the comparisons of the kNN and the SVM classifiers on the Letter Image Recognition Dataset and its reduced training set in terms of classification accuracy and the average CPU time, respectively. From Tables 2 and 3, we can see that although more than 55% of samples were removed, there weren't significant decreases in the success rates, and meanwhile, in the two classifiers, the time taken for the training (or classification) has remarkably reduced. For example, in the kNN classifier, the classification accuracy has decreased from 95.9 to 94.9 percent, while the time taken for classification has reduced from 33.4 seconds to 14.9 seconds.

TABLE II. PERFORMANCES OF THE *k*NN CLASSIFIERS

	Original set	Reduced set
Correctly classified (%)	95.9	94.9
Classification time (s)	33.4	14.9

TABLE III. PERFORMANCES OF THE SVM CLASSIFIERS

	Original set	Reduced set
Correctly classified (%)	93.5	91.1
Training time (h)	37.3	2.4

V. CONCLUSION

A method to overcome the problems with huge training datasets for classification is proposed. When the training set is huge, it is impossible to use the entire data set for training purposes due to space and computational limitations. Although some techniques have been developed for removing the irrelevant or unimportant samples in the training set, they were not designed for general proposal.

The presented algorithm is based on the Mean Shift procedure, which is a mode seeking procedure that estimates the gradient of the data density. We found that in a large number of samples to be trained, most samples can be clustered into a small number of mode centroids, thus, the original samples can be reduced though the use of a Mean Shift procedure. In conclusion, the algorithm is efficient because: (1) the number of removed samples can be controlled through modifying the parameter, with minimum effects on the distribution pattern; (2) unlike other clustering based sample selection methods, there is no need to manually identify the clusters.

In general, results obtained with the two datasets have shown that the proposed technique can be successfully applied to them, and the speed of training or classification can be decreased without significant effect upon the classification results. Although the classification rate is decreasing within an acceptable range, there is no guarantee of how much it will affect the final classification result. It is noteworthy that in some cases, the classification results can be improved when compared with using the whole training set. Therefore, further research will focus on maintaining the classification accuracy.

REFERENCES

[1] Fukanaga K., Hostetler L.D., "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans. on Information Theory, vol. 21, no. 1, pp. 32−40, 1975.

[2] Cheng Y., "Mean shift, mode seeking and clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, pp. 790−799, 1995.

[3] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift," in: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 142-149, 2000.

[4] D. Comanicu, P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, 2002.

[5] Parzen E., "On the estimation of a probability density function and mode," Annals of Mathematical Statistics, vol. 33, pp. 1065-1076, 1962.

[6] Keinosuke Fukunaga, "Introduction to Statistical Pattern Recognition 2nd Edition," Academic Press, New York, 1990.

[7] Jain A.K., Duin R.P.W., Mao Jian-Chang. "Statistical pattern recognition: A review," IEEE Trans. on Pattern Analysis Machine Intelligence, vol. 22, no. 1, pp. 4-37, 2000.

[8] Koggalage R., Halgamuge S., "Reducing the number of training samples for Fast Support Vector Machine Classification," Neural Information Processing Letters and Reviews, vol. 2, no. 3, pp. 57-65, 2004.

[9] Lyhyaoui, A. et al., "Sample selection via clustering to construct support vector-like classifiers," IEEE Trans. on Neural Networks, vol. 10, no. 6, pp. 1474-1480, 1999.

[10] Mykola Pechenizkiy, Seppo Puuronen, Alexey Tsymbal, "The impact of sample reduction on PCA-based feature extraction for supervised learning," In: Proc. of the 2006 ACM symposium on Applied computing, pp. 553-558, 2005.

[11] Turlach, B.A., "Bandwidth selection in kernel density estimation: A review," Discussion Paper. 9317, Institut de Statistique, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, 1993.

[12] Vojtˇech Franc, V´aclav Hlav´aˇ. Statistical Pattern Recognition Toolbox for Matlab, http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html.

[13] Frey P.W., Slate D.J., "Letter Recognition Using Holland-style Adaptive Classifiers," Machine Learning, vol. 6, no. 2, pp. 161-182, 1991.