Single-character OCR using Support Vector Machines

Olli Jarva & Jarno Rantanen

Aalto University School of Science

olli@jarva.fi & jarno@jrw.fi

Abstract

This paper describes a solution to an optical character recognition problem for bitmap characters using Support Vector Machines with an RBF kernel, including a description of RBF parameter search and bitmap normalization. Classification performance of 90.8% was achieved against a given training set of 42152 correctly labeled samples using k-fold cross validation with k=20 [1].

KEYWORDS: SVM, Support Vector Machine, RBF, OCR, Character Recognition

1 Data set description

The data set against which our solution was developed consisted of a provided set of 42152 black-and-white bitmaps, depicting hand-written instances of characters from the English alphabet (that is, the task was to classify the bitmaps into 26 distinct categories). Each bitmap was given as a 16-by-8 image, in the form of a binary vector of length 128 (meaning an array of 128 ones or zeroes). The vectors were delivered in a text file, with the correct label associated with each vector.

As an optional extra task, we were provided with the opportunity of participating in a competition amongst different solutions to the same classification problem. The competition was organized by first providing only a subset of the training data (10000 vectors), using that to train a classifier, and then calculating an error rate against the rest of the training data (which was not yet made available at that time). Our participation in the competition is discussed further in Section 8.

The same data set was used for both training our classifiers, and testing them afterwards. The data was split using k-fold cross-validation to minimize overfitting, while making the most of the available data. This testing technique is discussed in further detail in Section 6.

2 Method selection

There is no de-facto solution to the problem we were facing; Optical Character Recognition (OCR) is a wide field of research with ongoing investigation into new methods. An important factor in method selection was simplicity - in the beginning of the project, we were not seasoned experts in OCR or Machine Learning, so we were looking for established methods that work robustly even in the hands of beginners.

One initial contender was Principal Component Analysis (PCA). PCA works by reducing a high-dimensional data set into a (hopefully) smaller set of dimensions, so that the resulting dimensions capture most of the variance of the original space. PCA by itself is not a complete solution to our problem, as the data points would still have to be labeled in the dimensionally reduced input space. [4]

For this task, with or without the help of PCA, one could use basic clustering algorithms such as k-means clustering. k-means clustering works by identifying exactly k clusters from the input data, without supervision. We had pre-labeled training data at our disposal, however, so we could make use of supervised learning methods instead. [4]

Perhaps the simplest of all supervised machine learning methods is the k-nearest neighbor (kNN) algorithm. It works by assigning classes to samples by way of majority vote: for an incoming, unlabeled sample, the k nearest neighbors are looked up (using either simple Euclidean distance, or a custom distance function). The assigned class is decided simply by observing the class memberships of the neighbors and choosing the most common one. kNN is not guaranteed to work well with high-dimensional input data, however, and would likely require at least dimensionality reduction by PCA or other preprocessing methods. [8]

In the end we chose to go with Support Vector Machines (SVM's), which were suggested by our literature review as a relatively simple yet effective machine learning method for our type of classification problem (OCR). The underlying mathematics and theory of the method is beyond the scope of this paper, and ready-made implementations are available for several common programming languages. Thus, an intimate understanding of the inner workings of the algorithm would not be required, as we would only be configuring and using an SVM, and not implementing one from scratch. SVM's are described in more detail in Section 3 below. [6]

Based on the performance of our solution, we have reason to believe we made a sensible choice in using SVM's for this task. The performance aspects are discussed further in Section 8.

3 Support Vector Machines

Support Vector Machines (SVM's) are a method for supervised machine learning, meaning they solve a classification problem by creating a classifier function from a set of existing, labeled

data points. This function can then be used to classify (or *label*) subsequent data points *without* supervision. In contrast to *regression methods* which produce continuous output, the output of a classifier function is always exactly one class into which the input data point (likely) belongs.

In its most basic implementation, an SVM is trained with data labeled into exactly two separate groups. The resulting classifier is a *binary one*, meaning it will classify subsequent input into those same two categories. If the input data points belong to a two-dimensional space, this can be intuitively thought of separating the data points into two clusters. To minimize the potential for generalization error, the clusters should be separated by as wide a band as possible (or in simpler terms, by a line, with as much space between the line and the nearest data points as possible). [7]

In mathematical terms, an SVM solves an optimization problem for pairs $(x_i, y_i), i = 1, ..., l$ where x_i is the i:th input data vector, and y_i is the correct label associated with that input:

$$\min(w, b, \xi): \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i,$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where C is the penalty parameter (see Section 6 for how it is chosen), w is a support vector in the problem space, b is a scalar associated with the support vector, ξ is the error term, and ϕ is a mapping function for the input space (the role of which is described below). [6, 3]

SVM's generalize nicely into higher dimensional spaces. In an n-dimensional input space, the two classes are linearly separated by an (n-1)-dimensional hyperplane instead of a line. They also generalize into working with >2 classes by way of reducing the multi-class classification problem into a set of binary classification problems. This can be done, for example, by chaining the binary classifiers so that the first classifies the input data as either belonging to class "A" or "other", the second one to "B" or other, and so on. [5]

The above works off the assumption that the sets being discriminated are linearly separable in the input space. With realistic data sets, however, this is often not the case. To keep the sets linearly separable (and thus the SVM approach applicable), the input space can be mapped into a much higher-dimensional space. The assumption is that with this added sparsity, a linearly separating hyperplane can be found, even for problematic input data sets. [7]

Such mappings can be achieved using what are known as *kernel functions*. Kernel functions have special properties that, in addition to helping the linear separability, reduce the computational load bearable. Numerous kernel functions have been proposed in literature, and new ones are being researched. The performance characteristics of the functions are highly dependent on the type of classification problem at hand, and the properties of the input space, and not all kernels work for all problems. There is, however, little theoretical base on how to *choose* a suitable kernel for a given data set, and thus it is in fact common to simply rely on empirical methods and compare the performance some common kernels, choosing the one that best fits the specific data set. [7]

4 Kernel selection

The Radial Basis Function (RBF) kernel is a common first-choice kernel suggested by literature [6, 7]. It has several desirable properties. Firstly, it is numerically robust in comparison to some other kernels. Secondly, another common alternative - the linear kernel - is simply a special case of the RBF one. Thirdly, and perhaps most importantly, the SVM kernel is configured with only two parameters (as opposed to the 4 parameters of the polynomial kernel, for example). As the parameter search is in essence a local optimization problem, having a 2-dimensional search space makes this problem more manageable than, say, searching in 4 dimensions. The parameter search is discussed further in Section 6. [6]

We did briefly try out other kernel functions (as many are readily provided with our SVM implementation of choice, mlpy [2]), but we ended up agreeing with our first choice of RBF, as it seemed the best performer against our data set. Also, since a much bigger part of the effort in configuring an SVM has to do with choosing proper parameters for the kernel function, RBF had desirable properties in that regard as well. These are discussed further in Section 6.

5 Character preprocessing

An SVM operates on input data represented as vectors of real numbers. Scaling the input data is very important with SVM's; features of the input data with large numerical ranges can easily dominate ones with smaller ranges, even though the width of their range has no real correlation with their importance in the actual classification problem at hand. It is thus suggested to always scale the data into a normalized range of [-1,1], or even [0,1]. [6]

Much of the preprocessing for the input data in our experiment was already done for us; the image data was nicely encoded into a set of binary vectors, so no bitmap processing was required. Also, since the vectors were binary, no data scaling was required.

The single preprocessing technique we opted for was basic noise reduction, by moving each image to the bottom left corner. That is, making sure images otherwise identical except for their padding would still look identical to the algorithm. We found this to increase our initial classification performance by 0.5%.

6 RBF kernel parameter search

Optimal parameters for the SVM (and the kernel function) present a local optimization problem in a 2-dimensional search space. The dimensions being explored are γ (the RBF kernel parameter) and C (the SVM penalty parameter). Since an exhaustive search in this space isn't possible, a basic grid search was employed.

In the first step of the search, the space was divided into a grid of 6-by-6 (totaling 36 cells), at each of which the classifiers were trained and the performance of the algorithm was measured. Our literature review ([6] especially) suggested going for exponentially increasing values of the parameters γ and C when doing the coarse mapping of the search space, which seemed like a reasonable starting point. The coarse search step covered γ for 2^i with $i = -10, -8, -6, -4, -2, 0$ and C for 2^i with $i = -1, 1, 3, 5, 7, 9$. Out of the explored cells, the one at which the algorithm performed best was chosen for the next step.

In the second step, the chosen cell was further explored by fine-tuning the parameters around the cell origin. The search works by varying both parameters by a small percentage until the performance converges at a local maximum.

The aforementioned classification performance was measured using a technique known as *k-fold cross-validation*. In a naive implementation of classifier training, the input data would be divided into two different sets: a training set, which is used to train the classifier, and a validation set, that is used to measure the performance of the classifier (as both sets contain the correct labels for data points). This approach is not optimal, however: the classifier may suffer from *overfitting*, meaning it ends up modeling more the training set, instead of the actual classification problem being solved. This is due to the classifier being trained only against a specific part of the training data, and validated against another. How those sets are chosen (that is, picked from the entire set of labeled data available) can greatly affect the resulting classifier. Should the validation data set be chosen with a bit of bad luck, for example, it may end up containing a very specific subset of the entire data, skewing the resulting error rates.

K-fold cross-validation avoids these issues by dividing the labeled data into k-segments of equal size, and then using each as the validation set in turn. The rest of the segments are then used to train the classifier being validated. Once each segment has been validated against, the error rate is calculated over the entire set of validations. This both makes use of the entire data set for training, and avoids overfitting, as all parts of the data are (at one point) used for validation.

7 Classifier chaining

During initial testing we found that certain pairs of characters were often confused by the classifier. As an example, the characters "i" and "l" were commonly misclassified. To compensate for this, we adjusted the primary classifier to treat "i" and "l" (and similar other pairs of characters) as a single category. Then, after the primary classification, the input data points classified into these combined categories were sent to a second-level SVM, specifically trained to discriminate between the two characters in the combined category. This effectively turned our classifier into a *classifier tree*, with second-level classifiers chained to the primary SVM.

Even though this approach originally yielded classification performance improvements of

Character	Misclass. (%)	Misclass.
a	5.62%	217
b	1.76%	68
c	3.27%	126
d	2.20%	85
e	3.89%	150
f	3.94%	152
g	4.90%	189
h	3.03%	117
i	12.57%	485
j	1.43%	55
k	3.42%	132
l	5.21%	201
m	1.81%	70
n	5.47%	211
o	3.39%	131
p	1.58%	61
q	2.67%	103
r	5.78%	223
s	2.33%	90
t	3.96%	153
u	6.17%	238
v	4.72%	182
w	0.93%	36
x	1.92%	74
y	5.70%	220
z	2.33%	90

Table 1: Per-character misclassifications. Percentages are in relation to the total count of all characters misclassified by the SVM.

a few percent, after carefully selecting the primary classifier parameters γ and C, the original primary classifier performed better than the classifier tree. In the end, this was both desirable and intuitive. It was desirable because it makes our solution simpler, and simple solutions are easier to sanity-check. It was intuitive (especially in retrospect) because our solution was in fact analogous to how multi-class SVM's are often constructed internally anyway (as discussed in Section 3), so it would have been unexpected had our duplication of this construct yielded significant improvements.

8 Results and performance

We were happy with the performance of our solution. We achieved a classification performance of 90.8% against the given training set of 42152 correctly labeled samples, as measured by k-fold cross validation with k=20. On a 2.1GHz Xeon development machine, a full run of the training and validation phases (as the cross-validation requires several trainings and validations) for a given combination of γ and C took approximately 6 hours (single-threaded).

Per-character classification performance has been included in Table 1, and visualized in Figure 1. The most noteworthy dip in performance is for the character "i". It is a fairly common character in the data set (see Figure 2), so the performance drop shouldn't be an artifact of in-

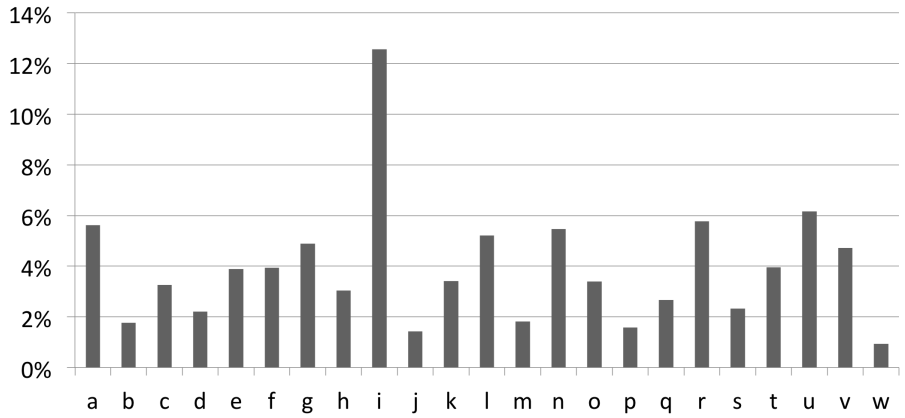

Figure 1: Per-character misclassifications, visualized. The higher the percentage, the more commonly the character was assigned with an incorrect label. See Table 1 for the source data.

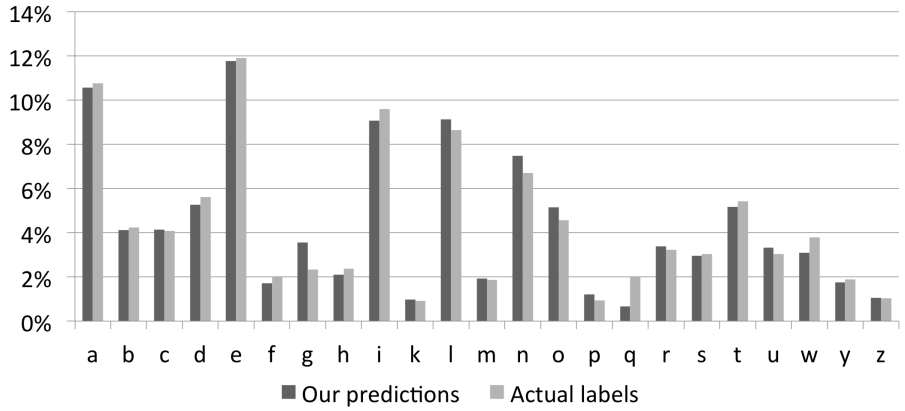

Figure 2: Character distribution in the data set. Our predictions (that is, labels assigned by our algorithm) have been contrasted with the correctly labeled data.

frequency and bad luck, for example. The visualization of pairwise misclassifications in Figure 3 helps explain this, as it shows there is another character (namely "l") which is *very* commonly confused with "i". Furthermore, Figure 4 highlights the visual similarity between the two.

From this map it is easy to recognize the most common pairs of characters to be misclassified. By far the most common pair is the aforementioned "i" and "l". This is understandable, considering the visual similarity of "i" and "l" as seen in Figure 4; the leftmost instance of "i" looks in fact exactly like an instance of "l". The second most common pair to be misclassified is "u" and "n", and the third most common pair is "n" and "a". The second "n" from the left demonstrates well the difficulty in discriminating them from instances of "u", whereas the first two instances of "a" (with their open tops) could very well be confused with an "n" or even a "u". Such a visual comparison is a good reminder as to how challenging the classification task at hand actually is, and helps appreciate the performance of the algorithm where even the human observer may occasionally fail.

It is interesting to note that the pairwise misclassifications are not symmetric, meaning that, for example, "i" is more commonly misclassified as an "l" than vice versa. Unfortunately, one

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a			0.9	2	1.4		0.6						0.4	3.8	2.6						1.3					0.9
b	0.2			0.2				0.2			0.1				0.2											
c	0.6				1.2				0.2			0.6			1.1			0.9			0.3					
d	0.5						0.2								0.6											
e	2.1		2.9												1.9	1.8		0.7		0.8						1.8
f					0.1											0.5		0.8	0.1	0.7						
g	1.2			0.7	0.3										0.7	0.3	0.5	0.6	2						2	
h		0.2									0.4	0.2		0.6							0.1					
i			0.8									41						1.3								
j				0			0		0.1										0							
k	0.1	0.2		0.2	0.3			0.5						0.3				0.3			0.3			0.2		
l			1		0.3				7.9												0.3					
m	0.4													1.3												
n	5						1.3				0.6		2		0.9			1			8.4					
o	2.9		0.6	0.6	0.8		2.4														2.2					
p					0.3		0.2											0.2							0.4	
q	0						0.3									0										
r	1.1		1.1		1.1	0.9			0.9		0.4	0.7		0.6		1.1				0.8		1.7			1.1	
s	0.2				0.2		1											0.2								
t				0.3	0.3	0.5		0.2	0.2		0.3	0.8		0.3				1.4							0.7	
u	2.5													3.7	1.2							3.6				
v															0.1			0.5			2					
w														0.1							0					
x											0.1							0.1							0.2	
y				0.3			3.4					0.2			0.2			0.3		0.3						
z	0.2				0.7		0.4													0.2						

Figure 3: Character misclassification map. The correct label can be read from the left, and the incorrect classification from the top. The values have been scaled with the relative frequency of the character in question. The highest 10% of values has been highlighted.

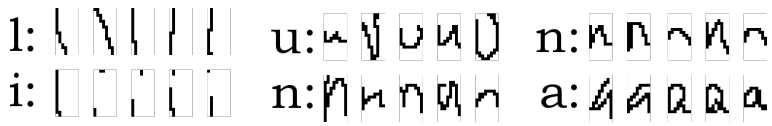

Figure 4: The most commonly misclassified character pairs, by example. Specimens have been randomly selected from the correctly labeled data set.

natural explanation of significantly differing frequencies does not check out, as the characters are approximately as frequent (see Figure 2). The cause for this (and potential improvements using this knowledge) remain an intriguing topic for further research.

The voluntary competition amongst other solutions to the same classification problem mentioned in Section 1 ended up favoring our solution. The closest contender (at an error rate of 12.08%, compared to our 10.48%) was, interestingly, an implementation of kNN, which was discussed as one of our early candidates in Section 2. What made the solution interesting was its simplicity - kNN is one of the simplest methods of machine learning, yet still quite effective with suitable kinds of problems [8]. As far as we know, the solution used a custom distance function instead of standard Euclidean distance, which makes it all the more interesting.

References

[1] OCR training set. http://users.ics.aalto.fi/kcho/MLBP2012/train.txt , retrieved at 18th Oct 2012. Instructions: https://noppa.aalto.fi/noppa/kurssi/t-61.3050/term_project.

[2] D. Albanese, R. Visintainer, S. Merler, S. Riccadonna, G. Jurman, and C. Furlanello. mlpy: Machine learning python, 2012.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[5] K. Duan and S. Keerthi. Which is the best multiclass svm method? an empirical study. *Multiple Classifier Systems*, pages 732–760, 2005.

[6] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classication, 2010. http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf , retrieved at 18th Oct 2012.

[7] O. Invanciuc. Applications of support vector machines in chemistry. *Reviews in Computational Chemistry*, 23:291–400, 2007.

[8] J. Keller, M. Gray, and J. Givens. A fuzzy k-nearest neighbor algorithm. *Systems, Man and Cybernetics, IEEE Transactions on*, (4):580–585, 1985.