IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 4, June 2014.
www.ijiset.com
ISSN 2348 – 7968

Handwritten Digit Recognition with Improved SVM

Ruta Ashok Kambli[1],Yogesh Kailas Ankurkar[2] and AmeyaVinay Mane[3]

[1] Electrical Dept., VJTI
Mumbai, Maharashtra, India

[2]Electrical Dept., VJTI
Mumbai, Maharashtra, India

[3]Electrical Dept., VJTI
Mumbai, Maharashtra, India

Abstract

In this paper Support Vector machine is used to recognize handwritten digits. Support Vector Machine is classification tool. This SVM is combined with the different dimension reduction techniques, to obtain better classification results. However, if assume the original data actually exists on a lower dimensional manifold embedded in a high dimensional feature space, then recently popularized approaches based in graph-theory and differential geometry allow us to learn the underlying manifold that generates the data. One such manifold-learning technique, called Diffusion Maps, is said to preserve the local proximity between data points by first constructing a representation for the underlying manifold. This work examines binary handwriting digit classification problems using Diffusion Maps to embed the data. Results show that diffusion map is well suited for this method than any other dimension reduction technique should summarize the content of the paper

Keywords:*Handwritten Digit Recognition, Support Vector Machines, Dimension Reduction, Diffusion Maps.*

1. Introduction

The procedure of transformation of images into understandable format which are hand-written, typewritten, or printed digits, for the purpose of editing, indexing/searching, and a storage size reduction, is commonly known as handwritten recognition. Handwritten recognition system has usefulness and importance in numerous fields such as processing bank check amount, recognizing the zip codes on mails for postal mail sorting, online data indexing, handwriting recognition on computer, numeric entries in the form filled by hand and so on. The handwritten recognition system consists of two distinct domains on the basis of input signals, online and offline.

The static representation of a digitized document is used in the offline system of digit recognition, example of which are check form, mail or document processing. Contrary to offline, in online system depends on the information acquired during the production of the handwriting. In the online handwritten recognition, the writing tool trajectory knowledge capturing instrument is required. New era electronics devices, such as smart phone, electronic pad and digital personal assistant have online handwritten recognition system in them. Hence it is important to improve and optimize the performance of the recognition system, aiming for the reducing the storage space and improving the processing speed. Figure 1 shows an online handwritten digit recognition system.

2. Support Vector Machines

Support Vector Machines [1], in the machine learning theory, are used for classification and regression

Fig. 1. Handwritten Digit Recognition

ISSN 2348 – 7968

analysis. They are supervised learning models with associated learning algorithms that analyse data and recognize patterns. As per the requirement of the data under consideration, the Support Vector Machines are modelled to classify.

The figure 1 shown below is flow chart of generalization of modelling of hand written digit recognition. First step is to model the SVM and then find the data set to train this model.

The hyperplane is generated during the training of the SVM model. This hyperplane classifies data into different classes. After formation of the hyperplane, offline data is gven as an input to the SVM model and classified output is checked for the correctness of its classification. The online data is given as an input to SVM model if offline data classification is optimized, if it is not, then kernel and other parameters are modified to obtained optimal classification. After modifying the parameters, SVM model is checked for offline data. If it is giving optimal solution, then online data is given as an input to the SVM.

2.1 Hard-margin Support Vector Machine

Let us consider that, M m-dimensional training inputs xi (i = 1, 2, . . , M) may belong to Class 1 or 2 and the associated labels be yi= 1 for Class 1 and -1 for Class 2. The decision function is defined for linearly separable data as:

$$y_i\left(w^T x_i + b\right) \geq 1 \; for \; i = 1, \ldots\ldots, M \qquad (1)$$

Where W is an m-dimensional vector, b is a bias term.

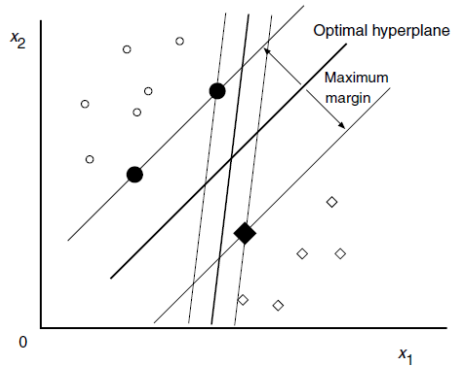

Fig. 2. Hard-margin SVM

Fig. 2 shows plot of two decision functions that satisfy equation (1). Thus this is clear from fig. 2 that , there may be infinite decision planes which can satisfies equation (1). These decision making planes are nothing but hyperplanes which seperates data into different classes. The generalization of SVM depends upon the location of separating hyperplane and the optimal separating hyperplane is one with maximum margin. The generalization ability is maximized if optimal separating hyperplane is same as the separating hyperplane, under the assumption that, no outliner will include in training data, and test data will follow the same distribution as of training data.Assume that no outliers are included in the training data and that unknown test data will obey the same distribution as that of the training data.

Therefore, the optimal separating hyperplane can be obtained by solving the following minimization problem for w and b:

Minimize :$Q(w,b) = \frac{1}{2}\|w\|^2$ (2)

Subjectto: $y_i(w^T x_i + b) \geq 1 \;\; for i = 1, \ldots\ldots, M$ (3)

In the hard margin SVM model, data is linearly severable. The data that satisfy the equalities are called support vectors [2]. In above figure the data corresponding to the filled circles and the filled square are support vectors. When data are linearly inseparable, there is no feasible solution and hard-margin SVM is unsolvable [3].

2.2 Soft-margin Support Vector Machine

To solve the problem of linearly inseparable data, the nonnegative slack variable ξ_i ($\geq$ 0) is introduced in equation (1) [4],

$$y_i(w^T x_i + b) \geq 1 - \xi_i \;\; for \;\; i = 1, \ldots\ldots, M \qquad (4)$$

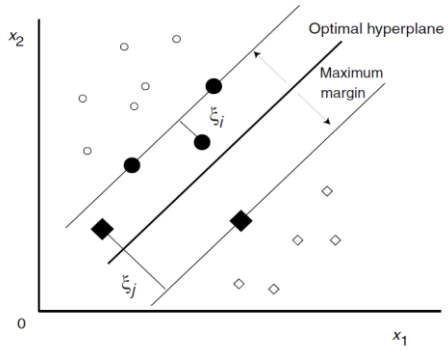

Fig. 3. Soft-margin SVM

Fig. 3 shows the soft margin SVM. By the slack variable ξ_i, feasible solutions always exist. For the training data xi, if 0 <ξ_i<1, the data do not have the maximum margin but are still correctly classified. But if $\xi_i \geq 1$ the data are misclassified by the optimal hyperplane. Now our minimization problem becomes;

$$minimise Q(w, b, \xi) = \frac{1}{2}\| w \|^2 + \frac{C}{P}\sum_{i=1}^{M} \xi_i^P \qquad (5)$$

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \; for i = 1, \ldots\ldots, M \qquad (6)$$

ISSN 2348 – 7968

Where $\xi = (\xi_{1,\ldots\ldots}\xi_M)^T$ and C is the margin parameter that determines the trade-off between the maximization of the margin and the minimization of the classification error. The value of P is selected as either 1 or 2. The obtained hyperplane is called the soft-margin hyperplane. When P = 1, the support vector machine the L1 called soft-margin support vector machine or the L1 support vector machine for short (L1 SVM). This paper used L1 SVM only.

2.3 Mapping to High Dimension space (Kernel trick)

In a support vector machine, to mazimize the generalization ability, a hyperplane is defined. But even if optimal hyperplane is defined, high generalization ability is not possible if data is non-linearly distributedhence to enhance the linear separability of the data, a high-dimensional dot-product space is defined and input data is mapped into that space. This space is called feature space. Now using the nonlinear vector function $\phi(x) = (\phi_1(x), \ldots\ldots, \phi_1(x))^T$ that maps the m-dimensional input vector xinto the l-dimensional feature space, the linear decision function in the feature space is given by

$$D(w^T \phi(x) + b)$$

Where w is an l-dimensional vector and b is a bias term.

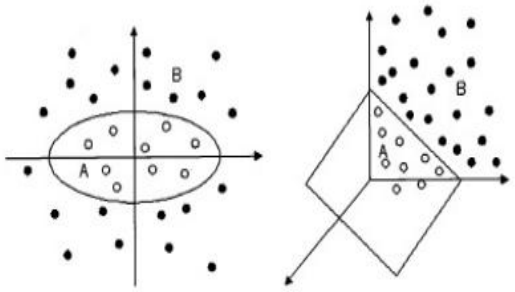

Fig. 4. Mapping to high dimension space

The mapping input data into feature space to avoid explicit treatment of variables in feature space by kernel tricks are called kernel methods or kernel based methods. [5]. By selecting proper kernel SVM can improve generalization performance, is important advantage of SVM. There are different types of kernels like linear kernel, polynomial kernel, radial basis function kernel, sigmoid kernel.

2.4Extension of SVM

2.4.1Multiclass SVM

Multiclass SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements[6]. The main reason to follow this approach is to reduce the single multiclass problem into multiple binary classification problems. Common methods for such reduction include:

1. Building binary classifiers which distinguish between (i) one of the labels and the rest (*one-versus-all*) or (ii) between every pair of classes (*one-versus-one*). Classification of new instances for the one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class (it is important that the output functions be calibrated to produce comparable scores). For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification[14].
2. Directes Acyclic Graph SVM (DAGSVM)[15].
3. Crammer and Singer proposed a multiclass SVM method which casts the multiclass classification problem into a single optimization problem, rather than decomposing it into multiple binary classification problems[16].

2.4.2 Regression

A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola[7]. This method is called support vector regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin[17]. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction (within a threshold) Another SVM version known as least squares support vector machine (LS-SVM) has been proposed by Suykens and Vandewalle[8].

3. Experiment

In the experiment, the problem of handwritten digit recognition is aimed. In this experiment, the handwritten digit data set available at UCI dataset is used [9].Each

ISSN 2348 – 7968

database is divided into ten groups that are as equal as possible, 10-fold cross validation. Nine groups are set aside for the training set and one group for the dedicated testing set. This procedure is continued until all groups have represented as a testing set. Diffusion map is applied to the data set. Basic diffusion maps algorithm is as follows[10]:

Algorithm 1 Basic Diffusion Mapping Algorithm

INPUT: High dimensional data set X_i, $i = 0, 1, 2 \ldots \ldots N-1$.

1. Define a kernel, $k(x, y)$and create a kernel matrix, *K*, such that $K_{i,j} = k(X_i, X_j)$
2. Create the diffusion matrix by normalizing the rows of the kernel matrix.
3. Calculate the eigenvectors of the diffusion matrix.
4. Map to the *d*-dimensional diffusion space at time *t*, using the *d* dominant eigenvectors and –values.

OUTPUT: For dimensional data set $Y_i, i = 0, 1 \ldots . . N-1$.

After that the data is classified using SVM. The accuracy of SVM classifier increases in case of diffusion process, as compare to other dimension reduction methods. The experiment shows how the algorithm integrates local information through a time-dependent diffusion to reveal structures at different time scales. The chosen data-set exhibits different structures on each scale.

The performance of diffusion mapping is compared to other dimension reduction techniques. The same data set is used to investigate the performance of other techniques. As the data under consideration is of nonlinear nature so the linear techniques such as principle Component Analysis (PCA)[13] and multidimensional scaling (MDS)[12] fail. According to theory of Isomap[11], it should work fine with nonlinear data, but it is less robust to the noisy data.

4. Conclusion

This paper investigated diffusion maps, a technique for nonlinear dimensionality reduction with Support Vector Machine, a non-probabilistic binary linear classifier. shod how it integrates local connectivity to recover parameters of change at different time scales. The method is compared to other three techniques and found that the diffusion mapping is more robust to noise perturbation, and is the only technique that allows geometric analysis at differing scales. Future work will revolve around applications in clustering, noise-reduction and feature extraction.

5. References

[1] N.Cristianini and J.Sha-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University press, 2000.

[2] Cortes, C. ,Vapnik, V. (1995). "Support-vector networks". Machine Learning 20 (3): 273. DOI :10.1007/BF00994018.

[3] V. David Sanchez A, "Advanced Support Vector Machines and Kernel Methods", Neuro Computing, vol.55, no.1, pp. 5-20, Sep, 2003.

[4] C.F. Lin, S.D. Wang, "Fuzzy Support Vector Machines", IEEE Trans on Neural Networks, vol.13, no.2, pp. 464-471, Mar, 2002.

[5] D.G. Chen, Q.He, X.Z.Wang, "FRSVMs: Fuzzy Rough Set Based Support Vector Machine", Fuzzy Sets and Systems, vol.161, no.4, pp. 596-607, Feb, 2010.

[6] J.H. Zhang, Y.Y.Wang, "A Rough Margin Based Support Vector Machine", Information Science, vol.178, no.9, pp. 2204-2214, May, 2008.

[7] C.F.Lin, S.D.Wang, "Training Algorithms for Fuzzy Support Vector Machines with Noisy Data", Pattern Recognition Letters, vol.25, no.14, pp. 1647-1656, Oct, 2004.

[8] Y.H.Qian, J.Y.Liang, C. Dang, "Consistency Measure, Inclusion Degree and Fuzzy Measure in Decision Tables", Fuzzy Sets and Systems, vol.159, no.18, pp. 2353-2377, Sep, 2008.

[9] A. Asuncion; D.J. Newman; (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[10] R. Coifman; S. Lafon, "Diffusion Maps," Applied and Computational Harmonic Analysis, special issue on diffusion maps and wavelets, vol. 21, pp. 5-30, July 2006.

[11] R. Coifman; S. Lafon; A. Lee; M. Maggioni; B. Nadler; F. Warner; S. Zucker, "Geometric Diffusions as a Tool for Harmonics Analysis and Structure Definition of Data: Multiscale Methods," Proc. Nat"l Academy of Sciences, vol. 102, no. 21, pp. 7432-7437, May 2005.

[12] S. Rois; L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, pp. 2323-2326, 2000.

[13] M. Belkin; P. Niyogi, "LaplacianEigenmaps for dimensionality reduction and data representation," Neural Computation, v.15 n.6, p.1373-1396, June 2003.

[14] Duan, K. B.; Keerthi, S. S. (2005). "Which Is the Best Multiclass SVM Method? An Empirical Study".Multiple Classifier Systems. Lecture Notes in Computer Science 3541. p. 278.doi:10.1007/11494683_28. ISBN 978-3-540-26306-7.

[15] Hsu, Chih-Wei; and Lin, Chih-Jen (2002). "A Comparison of Methods for Multiclass Support Vector Machines".IEEE Transactions on Neural Networks.

[16] Platt, John; Cristianini, N.; and Shawe-Taylor, J. (2000). "Large margin DAGs for multiclass classification".In Solla, Sara A.; Leen, Todd K.; and Müller, Klaus-Robert; eds. Advances in Neural Information Processing Systems.MIT Press. pp. 547–553.

[17] Dietterich, Thomas G.; and Bakiri, Ghulum; Bakiri (1995). "Solving Multiclass Learning Problems via Error-

ISSN 2348 – 7968

Correcting Output Codes".Journal of Artificial Intelligence Research, Vol. 2 2: 263–286.arXiv:cs/9501101. Bibcode:1995cs.....1101D.