Performance Comparison of SVM and *K*-NN for Oriya Character Recognition

Sanghamitra Mohanty, Himadri Nandini Das Bebartta
Department of Computer Science and Application
Utkal University, Vani Vihar
Bhubaneswar, India

***Abstract*—Image classification is one of the most important branch of Artificial intelligence; its application seems to be in a promising direction in the development of character recognition in Optical Character Recognition (OCR). Character recognition (CR) has been extensively studied in the last half century and progressed to the level, sufficient to produce technology driven applications. Now the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies. Researchers for the recognition of Indic Languages and scripts are comparatively less with other languages. There are lots of different machine learning algorithms used for image classification nowadays. In this paper, we discuss the characteristics of some classification methods such as Support Vector Machines (SVM) and *K*-Nearest Neighborhood (*K*-NN) that have been applied to Oriya characters. We will discuss the performance of each algorithm for character classification based on drawing their learning curve, selecting parameters and comparing their correct rate on different categories of Oriya characters. It has been observed that Support Vector Machines outperforms among both the classifiers.**

Keywords-Recognition; Features; Nearest Neighbors; Support Vectors.

I. INTRODUCTION

During the past thirty years, substantial research efforts have been devoted to character recognition that is used to translate human readable characters to machine-readable codes. Immense effort has been made on character recognition, as it provides a solution for processing large volumes of data automatically in a large variety of scientific and business applications. OCR deals with the recognition of optically processed characters rather than magnetically processed ones. OCR is a process of automatic recognition of characters by computers in optically scanned and digitized pages of text [3].OCR is one of the most fascinating and challenging areas of pattern recognition with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications [1] [2]. The term image classification refers to the labeling of images into one of a number of predefined categories. Although it seemed not a very difficult task for humans, it has proved to be a difficult problem for machines. Therefore, image classification challenge is very important for OCR. One of the important tasks in machine learning is the electronic reading of documents. All official documents, magazines and reports can be converted to electronic form using a high performance OCR. From past many years, many academic laboratories and companies are involved in research on printed recognition. The increase in accuracy of printed processing results from a combination of several elements i.e. the use of complex systems integrating several kinds of information., the choice of relevant application domains, and new technologies such as high quality high speed scanners and inexpensive powerful CPU's. In Character recognition system we required two things i.e. processing on data set and decision making algorithms. We can categorize preprocessing into three categories: the use of global transforms local comparisons and geometrical or topological characteristics. There are various kinds of decision methods have been used such as: various statistical methods, neural networks, structural matching and stochastic processing. Many recent methods developed by combining several techniques existing together in order to provide a better reliability to compensate the great variability of document processing.[4] To address this problem, designing and implementation of automatic image classification algorithms has been an important research field for decades. The methods popularly used in the early stage of OCR research and development are template matching and structural analysis [4]. The templates or prototypes in these early methods were simple design methods are insufficient to accommodate the shape variability of samples, and so, are not able to yield high recognition accuracies. For large sample data, the character recognition community has turned attention to classification methods *K*-NN and SVMs, are also actively studied and applied in pattern recognition.

In this paper, we discuss the strengths and weaknesses of classification methods that have been widely used, identify the needs of improved performance in character recognition, and suggest some research directions of classification that can help meet with these needs. We will focus on the classification of isolated (segmented) characters, though classification method.

II. CHALLENGES IN ORIYA CHARACTER SET

The Oriya script has descended from the Brahmi script sometime around the 11th century AD. Oriya is the official national language of India, most frequently used by common people in Orissa. Oriya alphabets consist of 268 symbols (13 vowels, 36 consonants, 10 digits and 210 conjuncts).

Figure 1. Oriya character set

In the above Fig. 1, it can be noted that out of 52 basic characters 37 characters have a convex shape at the upper part. The writing style in the script is from left to right. The concept of upper/lower case is absent in Oriya script. A consonant or vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character or conjuncts. Compound characters can be combinations of consonant and consonant, as well as consonant and vowel. As a consonant gets merged within another consonant to form a complex character, they lead to a lot of confusion and it gets difficult on the part of the programmer to classify them uniquely.

III. Overview of Classification Methods

We mainly discuss feature vector-based classification methods, which have prevailed structural methods, especially in printed character recognition. These methods include *K* - Nearest Neighborhood and Support Vector Machines. Comparison between k-NN and SVM method for speech emotion recognition has also been proposed by Muzaffar Khan et. al [5]. Amendolia et. al. has carried out a comparative study on k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening [6]. An Improvement on k-Nearest Neighbor Classification has been carried out by N. Suguna Using Genetic Algorithm [7].

After giving an overview of the classification methods, we will compare the performance of these popular methods applied on Oriya character sets. We summarize these classification methods in categories of statistical method and kernel method respectively.

A. Statistical method K-NN

K-Nearest Neighbor classifier is one among the instance-based method and it is also called as lazy algorithm. Statistical classifiers are rooted in the Bayes decision rule, and can be divided into parametric ones and non-parametric ones [8, 9]. Non-parametric methods, such as *k*-NN rule, are not practical for real-time applications since all training samples are stored and compared. Here the probabilities are unknown; the decision must be based on training samples from known classes. A simple and often deployed decision rule is the k nearest neighbor (*k*-NN) rule: decide for the class that is most frequent among the k nearest neighbors of the unknown pattern in the feature space. It is one of the simplest but widely using machine learning algorithms. In this, reviewing methods are applicable to distance based nearest neighbor classifiers. An object is classified by the "distance" from its neighbors, with the object being assigned to the class most common among its k distance-nearest neighbors. If k = 1, the algorithm simply becomes nearest neighbor algorithm and the object is classified to the class of its nearest neighbor. The computation time to test a pattern in *K*-NN depends upon the number of training samples m and the size of the feature vector n which is O (n*m).

Distance is a key word in this algorithm, each object in the space is represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance to calculate distance between two vector positions in the multidimensional space.

B. Kernel methods- SVM

Kernel methods, including support vector machines (SVMs) [10, 11] primarily and kernel principal component analysis (PCA), kernel Fisher discriminant analysis (FDA), etc., are receiving increasing attention and have shown superior performance in pattern recognition. An SVM is a binary classifier with discriminant function being the weighted combination of kernel functions over all training samples. After learning by quadratic programming (QP), the samples of non-zero weights are called support vectors (SVs). For multi-class classification, binary SVMs are combined in either one-against-others or one-against-one (pair wise) scheme [12]. Due to the high complexity of training and execution, SVM classifiers have been mostly applied. A strategy to alleviate the computation cost is to use a statistical or neural classifier for selecting two candidate classes, which are then discriminated by SVM [13]. Dong et al. used a one-against-others scheme for large set Chinese character recognition with fast training [14]. The SVM classifier with RBF kernel mostly gives the highest accuracy.

IV. Feature Extraction

The feature extraction is the integral part of any recognition system. The aim of feature extraction is to identify patterns by means of minimum number of features that are effective in discriminating pattern classes. Oriya language having a distinct visual appearance could be identified based on its discriminating features.

A. K-Nearest Neighbor

A *K*-NN classifier in its most basic form operates under the implicit assumption that all features are of equal value as far as the classification problem is concerned. When irrelevant and noisy features influence the neighborhood search to the same degree as highly relevant features, the accuracy of the model is likely to lose. *Feature weighting* is a technique used to approximate the optimal degree of influence of individual features using a training set. When successfully applied relevant features are attributed a high weight value, whereas irrelevant features are given a weight value close to zero. *Feature weighting* is used in our experiment not only to improve classification accuracy but also to discard features with weights below a certain threshold value and thereby increase the efficiency of the classifier. We here consider of Feature weighting on basis of *longest- run features* with respect to wrapper-based feature weighting algorithm. This helps to find the globally optimal feature weights by means of a greedy local search.

Longest-run Features: For computing longest-run features from a character image, the minimum square enclosing the image is divided into 25 rectangular regions. In each region, 4 longest-run features are computed row wise, column wise and along of its major diagonals. The row wise longest-run feature is computed by considering the sum of the lengths of the longest run bars that fit consecutive black pixels along each of all the rows of a rectangular region, as illustrated. The three other longest-run features are computed in the same way but along all column wise and two major diagonal wise directions within the rectangular separately. Thus in all, 25x4=100 longest-run features are computed from each character image.

B. Support Vector Machines

The objective of any machine capable of learning is to achieve good generalization performance, given a finite amount of training data, by striking a balance between the goodness of fit attained on a given training dataset and the ability of the machine to achieve error-free recognition on other datasets. With this concept as the basis, support vector machines have proved to achieve good generalization performance with no prior knowledge of the data. The principle of an SVM is to map the input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyper plane with maximum margin between the two classes in the feature space[15].

An SVM in its elementary form can be used for binary classification. It may, however, be extended to multi class problems using the one-against-the-rest approach or by using the one-against-one approach. We begin our experiment with SVM's that use the Linear Kernel because they are simple and can be computed quickly. For a two class support vector machine let m-dimensional inputs xi (i = 1. . . M) belongs to Class 1 or 2 and the associated labels be yi = 1 for Class 1 and −1 for Class 2. Let the decision function be

$$D(x) = W^t x + b \quad (1)$$

where w is an m-dimensional vector, b is a scalar, and

$$y_i D(x_i) \geq 1 - \xi i \quad \text{for} \quad i = 1, \ldots, M \quad (2)$$

Here ξi are nonnegative slack variables. The distance between the separating hyper plane D(x) = 0 and the training datum, with ξi = 0, nearest to the hyper plane is called margin. The hyper plane D(x) = 0 with the maximum margin is called optimal separating hyper plane. To determine the optimal separating hyper plane, we minimize

$$\tfrac{1}{2}||w||^2 + C \sum_{i=1}^{M} \xi i \quad (3)$$

subject to the constraints:

$$y_i (w^t x_i + b) \geq 1 - \xi i \quad \text{for} \quad i = 1, \ldots, M \quad (4)$$

where, C is the margin parameter that determines the tradeoff between the maximization of the margin and minimization of the classification error. The data that satisfy the equality in (4) are called support vectors.

A priori, one does not know which value of soft margin will yield the classifier with the best generalization ability. We optimize this choice for best performance on the selection portion of our data.

V. Performance Comparision and Evaluation

The experiments of character recognition reported in the literature vary in many factors such as the sample data, pre-processing technique, feature representation, classifier structure and learning algorithm. A better scheme to compare classifiers is to train them on a common feature representation.

A. Kernel method Vs statistical method

For Kernel methods like SVMs, the parameters of one class are trained on the samples of all classes. For statistical classifiers, the parameters of one class are estimated from the samples of its own class only. We compare the characteristics of two kinds of classifiers in the following respects.

1) *Complexity of training:* The parameters of *K*-NN classifiers are generally adjusted by distance measure. By feeding the training samples a fixed number of sweeps, the training time is linear with the number of samples. SVMs are trained by quadratic programming (QP), and the training time is generally proportional to the square of number of samples.

2) *Flexibility of training*: The parameters of *K*-NN classifiers can be adjusted in Feature weighting improve classification accuracy for global performance and also easy to add a new class to an existing classifier. On the other hand, SVMs can only be trained at the level of holistic patterns. This classifier is proportional to square of the number of classes, and to guarantee the stability of parameters, adding new classes or new samples need re-training with all samples.

3) *Classification accuracy:* SVMs have been demonstrated superior classification accuracies to *K*-NN classifiers in many experiments. When training with enough samples, SVM classifiers give higher accuracies than statistical classifiers.

4) *Complexity of training storage and executon complexity:* At same level of classification accuracy, SVM learning by Quadratic Programming often results in a large number of SVs, which should be stored and computed in classification. *K*-NN classifiers have much less parameters, and are easy to control. In a word, *K*-NN classifiers consume less storage and computation than SVMs.

B. Evaluation

We tested the performance on Oriya characters. As till date no standard dataset is available for Oriya Characters we have collected and created the dataset within our organization. For this a corpus for Oriya OCR consisting of data base for machine printed Oriya characters has been developed. Mainly samples have been gathered from laser print documents, books and newspaper containing variable font style and sizes. A scanning resolution of 300 dpi is employed for digitization of all documents. The training and testing set comprises of 10, 000 samples each. Fig. 2. Shows some sample characters of various fonts of Oriya script used in the experiment.

We have performed experiments with different types of images such as normal, bold, thin, small, big, etc. having varied sizes of Oriya characters. We have considered gray scale images for collection of the samples. This database can be utilized for the purpose of document analysis, recognition, and examination. The training set consists of binary images of 297 Oriya letters. We have kept the same data file for testing and training for all types of different classifiers to analyze the result.

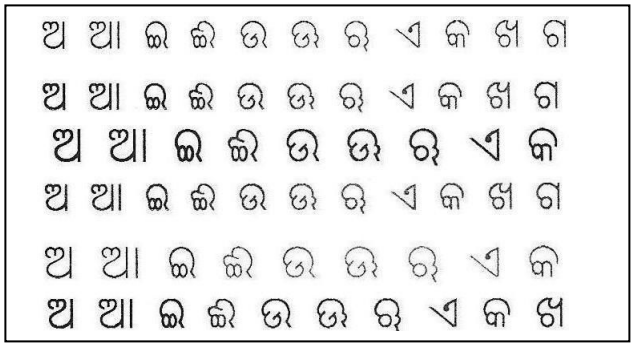


Figure 2: Samples of machine printed Oriya characters used for training

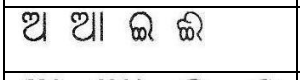
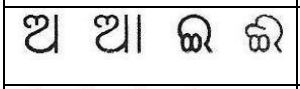
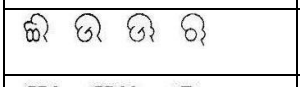
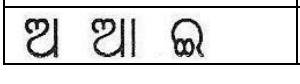
Below in Table 1 we can see the accuracy rate that we have obtained after testing the data set.

TABLE I. CLASSIFIERS ACCURACIES ON ORIYA CHARACTERS

Classifier	Training set	Test set	Accuracy Rate (in %)
SVM	10,000	10,000	98.9
KNN	10,000	10,000	96.47

Regarding the effect on accuracy by considering the different classifiers with different types of the images used for characters, for Oriya-Bold and big characters the accuracy rate is high in case of support vector machines and it has nearly 98.9 percentage of accuracy. The accuracy rate decreases for the thin and small size characters. Table 2 shows the effect on accuracy for Oriya by considering different character sizes with different types of the images using Support Vector Machines.

TABLE II: EFFECT ON ACCURACY BY CONSIDERING DIFFERENT CHARACTER SIZES WITH DIFFERENT TYPES OF THE IMAGES USED FOR ORIYA CHARACTERS.

Image type	Size of the samples	Accuracy percentage
	Bold and small	92.78%
	Bold and big	98.9%
	Normal and small	96.98%
	Normal and Bold	97.12%

VI. CONCLUSION

The results obtained for recognition of Oriya characters are quite encouraging and show that reliable classification is possible using SVMs. Nearest neighbor classifiers consume less storage and computation than that of SVMs. Biggest limitation of the support vector approach lies in the choice of kernel. Second limitation is in speed and size, both in training and testing. We applied SVMs and *K*-NNs classifiers on same feature data. Further we will find out the accuracy rate by taking a different set of samples for the test set and this work in future can be extended with degraded, noisy machine printed and italic text.

ACKNOWLEDGMENT

We are thankful to DIT, MCIT and my colleague Mr. Tarun Kumar Behera for their help and support.

REFERENCES

[1] J. Mantas, "An overview of character recognition methodologies", Pattern Recognition, vol. 19, pp. 425-430, 1986.

[2] V. K. Govindan, and A. P. Shivaprasad, "Character recognition –a review", Pattern Recognition, vol. 23, pp. 671-683, 1990.

[3] Pal, U., and B. B. Chaudhuri, "Indian script character recognition: a survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[4] S. Mori, C.Y. Suen, and K.Yamamoto, "Historical review of OCR research and development", Proceedings of the IEEE, vol. 80, pp. 1029-1058, July 1992.

[5] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin , and Ruhina Quazi, "Comparison between k-nn and svm method for speech emotion recognition", International Journal on Computer Science and Engineering, vol. 3, pp. 607- 611, February 2011.

[6] Amendolia et. al., "A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening" Chemometrics and Intelligent Laboratory Systems, vol. 69, pp. 13-20, August 2009.

[7] N.Suguna1, and Dr. K. Thanukodia, "An improved k-nearest neighbor classification using genetic algorithm", IJCSI International Journal of Computer Science Issues, vol. 7, pp. 18-21, July 2010

[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed, Academic Press, 1990.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed, Wiley Interscience, 2001.

[10] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New Work, 1995.

[11] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition, Knowledge Discovery and Data-Mining, 1998, pp. 1-43.

[12] U. Kressel, Pairwise Classification and Support Vector Machines, 1999, pp. 255-268.

[13] A. Bellili, M. Gilloux, and P. Gallinari, "An MLP-SVM combination architecture for online handwritten digit recognition: reduction of recognition errors by support vector machines rejection mechanisms", International Journal of Document Analysis and Recognition, vol. 5, pp. 244-252, 2003.

[14] J.X. Dong, A. Krzyzak, C.Y. Suen, "High accuracy handwritten chinese character recognition using support vector machine", Proceedings of the International Workshop on Artificial Neural Networks for Pattern Recognition, 2003.

[15] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998, pp. 121-167.

AUTHORS PROFILE

Prof. (Dr.)Sanghamitra Mohanty is the Professor and Head of the Department of Computer Science and Application, Utkal Unversity, Bhubaneswar,. Odisha, India. Her research interests are Speech processing, Image processing and Natural Language Processing. She has published 132 papers in International Journals and Conferences. She has guided many Ph. D. students and has 10 IPRs to her credit.

Himadri nandini Das Bebartta is continuing her PhD under the guidance of Prof. (Dr.) Sanghamitra Mohanty in the Department of Computer Science and Application, Utkal Unversity, Bhubaneswar, Odisha, India. She has been a research scholar since April 2007. And has worked as a Junior research fellow in the project development of robust document analysis and Recognition system for printed Indian scripts, DIT, MCIT, India.