Indian Journal of Science and Technology, *Vol 7(1), 35–38, January 2014*

Naïve Bayes Classifier with LU Factorization for Recognition of Handwritten Odia Numerals

Pradeepta K. Sarangi[1*], P. Ahmed[2] and Kiran K. Ravulakollu[2]

[1]Apeejay Institute of Technology, Greater Noida, U.P India; pradeepta_sarangi@yahoo.com
[2]Sharda University, Greater Noida, U.P, India; kiran.ravulakollu@sharda.ac.in

Abstract

We report the effectiveness of LU factorization (also called LU decomposition) as a technique for feature extraction along with naïve Bayes Classifier in recognizing handwritten Odia numerals (a regional language of north-eastern states of India derived from the Devanagari script). Experimental results show that LU factorization could be an alternative choice for feature extraction in pattern classification problems.

Keywords: Handwritten Odia Numerals, LU Factorization, Naïve Bayes Classifier.

1. Introduction

Recognition of handwritten scripts has been an area of intensive research. The wide scope of this field has attracted many national and international researchers. Researchers to find suitable techniques for feature extraction have put a lot of efforts. However, performance is far from perfection.

Feature extraction refers to representation of object in miniature preserving its originality. This reduction in dimensions helps in rapid efficiency solving complex issues, using computer programming with a less number of memory requirements. In pattern analysis, no standard universal feature extraction technique is available. This has to be developed, based on the nature of the script or pattern. Fewer numbers of elements in the feature vector may not be able to represent properties of the original object. On the other hand, too many elements in the feature vector may not meet the basic objective. Hence, the number of elements in the feature vector should be selected carefully.

Research in Odia language has not been much explored for over a long period. The condition for handwritten Odia script recognition is still worse. In India, few research centers are involved in exploring Odia scripts. Utkal University, Bhubaneswar is the only center in Odisha engaged in research on Odia scripts[1]. Here are presented some of the recent works on handwritten Odia numerals.

Sarangi et al.[2] proposes a feature extraction technique based on LU factorization of the image as a character matrix. The authors have used neural network as the classifier. Handwritten numerals data set generated by Indian Statistical Institute (ISI), Kolkata, India has been used for the experimental works with an accuracy of 85.30% reported.

Another work by Sarangi et al.[3] describes the use of Hop Field network as classifier with binary image as the feature vector.

Here, the authors have used their own collected handwritten numerals with reported accuracy of 95.4%. Some of the recent works on handwritten Odia numerals where the authors have addressed various feature extraction techniques and classifiers are summarized in Table 1.

Table 1. Summary of recent works

Authors	Features	Classifier
Sarangi et al.[2]	LU factorization	Neural Network
Sarangi et al.[3]	Binary Image	Hop Field
Pal et al.[4]	Directional	MQC
Roy et al.[5]	Directional	Quadratic
Bhowmik et al.[6]	Scalar	HMM

**Author for correspondence:*

In this paper, an attempt has been made to study and analyze the suitability of LU factorization as a technique for feature extraction. The naïve Bayes model has been used as a classifier to classify handwritten Odia numerals (an Indian regional script).

2. Naive Bayes Classifier

A naive Bayes classifier (also known as Bayesian Classifier) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independent assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[7].

Given B_1, B_2, …, B_N, a partition of the sample space S. Suppose that event A occurs; then using the definition of conditional probability and the theorem of total probability we obtain the probability of event B_j[8]:

$$P\left[B_j|A\right]=\frac{P[A\cap B_j]}{P[A]}=\frac{P[A|B_j].P[B_j]}{\sum_{k=1}^{N}P[A|B_k].P[B_k]} \quad (1)$$

For pattern recognition, Bayes Theorem can be expressed as:

$$P(\omega_j|x)=\frac{P(x|\omega_j).P(\omega_j)}{\sum_{k=1}^{N}P(x|\omega_k).P(\omega_k)}=\left[\frac{P(x|\omega_j).P(\omega_j)}{P(x)}\right] \quad (2)$$

where,

$P(\omega_i)$ = Prior Probability of Class ω_i.

$P(\omega_i|x)$ = Posterior Probability of class ω_i given the observation x.

$P(x|\omega_i)$ = Likelihood (conditional probability of x given class ω_i).

$P(x)$ = A normalization constant.

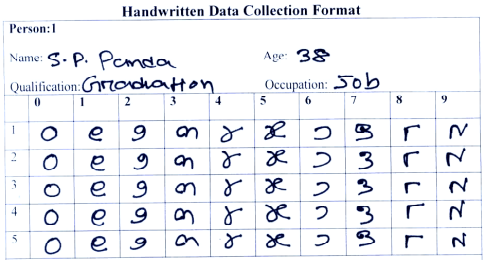

Figure 1. Sample format for data collection.

3. Database

No knowledgeable standard database on handwritten Odia numerals is available in public domain. However, ISI Kolkata has developed a database consisting of 5970 handwritten Odia numerals. This research has used the aforesaid database provided by ISI, Kolkata. In addition, we have also created a small database on handwritten Odia numerals consisting of 1300 numerals making a total sample size of 7270 for this experiment. A sample data collection format used by the authors is given in Figure 1.

4. Feature Extraction

In our earlier work[2], we introduced the use of LU factorization as a feature extraction technique. This research has made an attempt to explore more on the suitability of LU factorization as a technique for extraction of feature vectors. The key idea behind using LU factorization is to reduce the dimensionality of the original character image (generally in a higher dimension) to a lower dimensional space keeping the properties of the character image intact so that fewer memory space could be used to handle large number of character images.

A m×n matrix is said to have a LU factorization if there exists matrices L and U with the following properties:

(i) L is a m×n lower triangular matrix with all diagonal entries being 1.

(ii) U is a m×n matrix in some echelon form.

(iii) $A = LU$.

If $A = L_1U_1 = L_2U_2$ are two LU factorisation of a non-singular A, then

$$L_2^{-1}L^1 = U_2U_1^{-1} \quad (3)$$

Since the left part of the equation is a unit lower triangular matrix while the right side is an upper triangular matrix, both of the matrices must be the identity to satisfy the equation. Hence, $L_1 = L_2$ and $U_1 = U_2$.[9]

The following procedure has been adopted for extraction of L and U factors.

Let A be a square matrix. Then the LU factorization of A is of the form $A = LU$ (4)

Where L is a lower triangular matrix and U is an upper triangular matrix. This means that L has only zeros above the diagonal and U has only zeros below the diagonal.

Now consider a linear equation $AX = B$ (5)

Here 'A' is the character image in the matrix form. We have to only find out the value of 'X' using some values of 'B'. This can be achieved by using the formula:

$$X1 = B\backslash L \quad (6)$$

$$X2 = B\backslash U \quad (7)$$

where, X1 and X2 are the sets of feature vectors and L & U are the LU factorization of the image matrix. B is a column matrix.

5. Experimental Design

Since, LU factorization produces two sets of feature vectors; hence it was decided to implement all factors (L, U and LU combined). The aim was to test and analyze the scope and applicability of the technique in a broader sense.

A total of five different experiments have been carried out during this research. The experiments are categorized on the basis of feature vectors (i.e both L & U factors, L factors and U factors) and sample size as detailed in Table 2.

6. Results Analysis and Discussion

We studied the applicability of LU factorization as a technique for feature extraction. Different experiments have been carried out with different data sets and different factors of LU factorization. Naïve Bayes network has been used as a classifier. Results from different experiments have been summarized in Table 3.

Table 3 shows that the experimental results obtained using both parts of LU factorization are not very encouraging. The highest and lowest percentages of accuracy obtained are 85% for the numeral 9 and 74.39% for the numeral 7. The possibility of the lowest accuracy for the class 7 could be due to its similarity with the class 2.

In the second and third experiments, the L and U factors have been used separately as feature vectors. The objective was to analyze the factors independently. Here, we observed that the U factors are more effective than L factors. However, we can only conclude this from our experimental results.

The fourth and fifth experiments, where we have used a smaller size of data set, produce better results in comparison with previous experiments. The results are more encouraging as we have achieved an accuracy of 92.75%. Results from similar other works are summarized in Table 4.

The authors of previous reports[4,5] have used directional features and reported higher accuracy rate. However, the authors[6] have used scalar feature and reported a comparatively low accuracy (Table 4).

Table 2. Summary of experimental design

Experiment	Feature Vector	Sample Size
Experiment-1 (LU-4800)	Both 'L' and 'U' factors	4800
Experiment-2 (L-2400)	Only 'L' factors	2400
Experiment-3 (U-2400)	Only 'U' factors	2400
Experiment-4 (L-1200)	Only 'L' factors	1200
Experiment-5 (U-1200)	Only 'U' factors	1200

7. Conclusion

LU factorization could be a useful technique for extracting feature vectors in case of handwritten Odia scripts. Both the L and U factors have shown promising results. However, in our case the U factor has been proved as a better choice. Naïve Bayes classifier with LU factorization could be a good choice for recognition of handwritten Odia scripts. The combination of Naïve Bayes classifier with LU factorization performs better on a smaller size of data set.

Table 3. Summary of results from all experiments

Class	LU4800	L2400	U2400	L1200	U1200
0	80.60	87.91	88.33	87.5	90.83
1	83.93	87.5	88.75	86.66	90.83
2	78.33	88.33	87.91	90.83	92.5
3	76.21	87.08	87.5	90.83	93.33
4	81.96	83.75	88.33	93.33	94.16
5	78.63	87.91	88.75	91.66	92.5
6	82.27	87.91	86.66	92.5	94.16
7	74.39	85.83	86.25	90	93.33
8	84.24	82.08	88.33	91.66	90.83
9	85	88.33	87.91	92.5	95
N. A	80.56	86.66	87.87	90.75	92.75

N.A: Network Accuracy

Table 4. Results from similar works

Authors	Classifier	Features	Accuracy (%)
Sarangi et al.[2]	Neural Network	LU factorization	85.30
Sarangi et al.[3]	Hop Field	Binary Image	95.4
Pal et al.[4]	MQC	Directional	98.40
Roy et al.[5]	Quadratic	Directional	94.81
Bhowmik et al.[6]	HMM	Scalar	90.50

8. Recommendations for Future Work

The applicability of LU factorization as a feature extraction technique needs to be tested with other scripts. Combination of other classifiers with LU factorization needs to be tested for improved accuracy.

9. Acknowledgement

The authors are highly thankful to Prof. U. Pal and Prof. Ujjwal Bhattacharya (ISI, Kolkata) for providing us valuable research articles and Odia numerals database.

10. References

1. Pal U, Jayadevan R, Sharma N. Handwriting recognition in indian regional scripts: a survey of offline techniques. ACM Transactions on Asian Language Information Processing. 2012; 11(1):1–36.
2. Sarangi PK, Ahmed P. Recognition of handwritten odia numerals using artificial intelligence techniques. Int J Comput Sci Appl. 2013; 2(2):41–48.
3. Sarangi PK, Sahoo AK, Ahmed P. Recognition of Isolated handwritten numerals using hopfield neural network. Int J Comput Appl. 2012; 40(8):36–42.
4. Pal U, Wakabayashi T, Sharma N, Kimura F. Handwritten numeral recognition of six popular Indian scripts. Proceedings of 9th ICDAR; 2007 Sept 23–26; Parana. p. 749–753.
5. Roy K, Pal T, Pal U, Kimura F. Oriya handwritten numeral recognition system. Proceedings of ICDAR; 2005 29 Aug-1 Sept; 2:770–774.
6. Bhowmik TK, Parui SK, Bhattacharya U, Shaw B. An HMM based recognition scheme for handwritten Oriya numerals. Proceedings of 9th ICIT; 2006 Dec 18–21; Bhubaneswar. 105–110.
7. Zhang S, Gu M. improved text classification technique to acquire job opportunities for disabled persons. Computational Intelligence and Intelligent Systems, 5th International Symposium, ISICA, Wuhan, China; 2010. p. 280–287.
8. Gutierrez-Osuna R. Introduction to pattern analysis [Lecture notes]. Wright State University; notes provided at lecture given 2000, Available from: research.cs.tamu.edu/prism/lectures/iss/iss_l9.pdf.
9. Yang M. Matrix decomposition. Northwestern University. Evanston. [cited 2013 May 25]. Available from: http://users.eecs.northwestern.edu/~mya671/files/Matrix_YM_.pdf