Ensemble of Weak Learners on Recognition of Odia Numeric Digits

Pushpalata Pujari
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

Babita Majhi
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

***Abstract*- There are several applications of character recognition system in rural area like form processing, postal information processing, rural bank information processing, application processing. Most of the data collected from rural peoples are in regional language and in handwritten format. Recognition of such data is of vital importance in present scenario. This paper presents an ensemble model for recognition of Odia numeric digits. For the ensemble model a number of classifiers like Bayesian network, QUEST (Quick Unbiased and Efficient Statistical Tree), C5.0 and CHAID (Chi-squared Automatic Interaction Detector) are considered. The dataset consisting of 4000 Odia numeric digits is collected from ISI Calcutta. First the dataset is preprocessed and features are extracted by using gradient based approach. Then feature is reduced by using PCA (Principal Component analysis). The dataset with reduced feature is applied on weak learners and their outcomes are combined by using confidential weighted voting scheme to generate a number of ensemble model. Further the outcome of the ensemble model is combined with other weak learners to form new ensemble model. A number of ensemble models are constructed with various combinations of weak learners. It is found that the ensemble model constructed in succession yields greater accuracy than the other models. Thus a number of weak learners can be combined to form an ensemble model to sufficiently increase the classification accuracy of a recognition system.**

Keywords- Bayesian network; QUEST; C5.0; CHAID; Ensemble model

I. INTRODUCTION

A large number of populations of our country reside in rural areas. Several initiative schemes are being carried out by various organizations for the betterment of rural mass .Every scheme needs collection of large number of data from the rural areas. Most of the data collected from rural individuals are in the form of handwritten regional language. Often these data exist on paper and they have to be typed into the computer by human operators, for example, billions of letters, checks, payment slips, tax forms and application forms. Recognition of such data is of vital importance in present scenarios. Character recognition is an important technology for identification of handwritten characters written in regional language. Character recognition [1] is the conversion of printed or handwritten characters into machine readable form. In this paper a system is suggested for recognition of Odia numeric digits (0-9). Figure 1 shows the ten Odia numeric digits with their corresponding English numeral. Figure 2 shows ten Odia handwritten numerals from 0-9. The data set is collected from ISI Calcutta. The data set consists of 4000 samples of Odia handwritten numeric digits. Each digit (0-9) appears 400 times in the dataset. Basic steps involved in character recognition are preprocessing, feature extraction and classification. In this paper first the data set is preprocessed and feature extraction techniques based on gradient approach is applied on the dataset to extract features. Then obtained features are reduced by using PCA technique. After feature reduction recognition step is carried. The dataset is applied on various classifiers Bayesian network, QUEST, C5.0 and CHAID. Their outcomes are used for creating a number of ensemble models to improve the recognition rate.

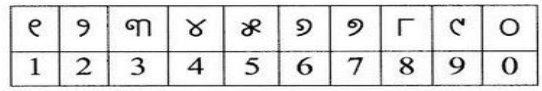

Figure 1 Ten Odia numerals with corresponding English numerals

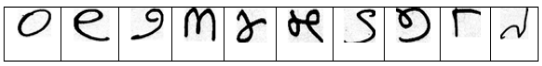

Figure 2 Handwritten Odia numeric Digits

The paper is organized as follows. Section II discusses the literature review. Section III discusses the proposed technology which includes preprocessing, Section IV describes feature extraction and feature reduction process. Section V describes classification and ensemble model. Section VI describes the simulation and experimental results. The result section is followed by conclusion and future scope of work.

II. LITERATURE REVIEW

From literature a few work has been done on ensemble of weak learners in the field of character recognition. Some of the work carried out is as follows.

I. Czarnowski and P. Zedrzejowicz proposed and validated an ensemble model to mine data streams with concept-drift constructed from the one-class base classifiers in [2]. In [3] the authors have proposed an ensemble model for improving the classification accuracy of NN (Nearest Neighborhood) Classifier. They used different distance function and different set of features for each classifier. The proposed model was tested on various UCI machine learning dataset. An ensemble model was proposed by Koichiro

Yamauchi [4] by using (RBFNNS) Radial Basis Function Neural Network to achieve model selection incrementally under virtual concept drifting environment where the distribution of learning samples varies over times. In [5] an ensemble classifier method is discussed with boosting, Adaboosting, Random Forest and general boosting projection method for fault diagnosis of wind turbines. The proposed method achieved highest accuracy with Adaboost using C5.0 decision tree as base classifier. In [6] each individual classifier was trained on a particular writing style. The proposed ensemble method was applied on two large scale handwritten word recognition tasks. An ensemble model was applied for recognition of offline cursive handwritten word in [7]. Hidden Markov models (HMMs) were used for the recognition.

From literature review the weak learners are ignored in the recognition process because of their slow learning rate. Hence an attempt is made in this paper to improve the recognition accuracy by using ensemble model constructed form the outcomes of weak learners.

III. PREPROCESSING

The method of extracting text from the document is called preprocessing. Preprocessing involves a series of steps like noise reduction, binarization, normalization, skew detection and thinning etc. The sample consisting of 4000 numeric digits is normalized into a standard pixel size 64x64. Normalization is carried out to remove variants in images. For obtaining gray scale image min filtering method is applied. Figure 3 represents the gray scale image of the numerals.

Figure 3 Gray Scale Images of the numerals

IV. FEATURE EXTRACTION

Feature extraction process is mainly used to extract important characteristics or feature in the images so that by using the extracted features it is possible to classify the numerals with less complexity. Once the important features are extracted it can be reduced further. In this paper gradient based approach is used for feature extraction. Robert filter [8], [9] is applied for obtaining gradient features. Figure 4 and 5 shows image of the direction of gradient and strength of gradient respectively. After the generation of features PCA is applied to reduce the features from 2519 to 75 numbers.

Figure 4 Image showing Direction of Gradient

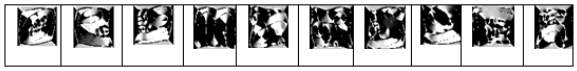

Figure 5 Image showing Strength of Gradient

V. CLASSIFICATION

The purpose of classification [10], [11] is to assign each data point with a class level. For classification the whole data set is divided into 90% of training and 10% of testing dataset. The preprocessing, feature extraction and feature reduction steps are carried out before classification. The dataset comprising of 3600 samples with reduced features are applied to the individual classifiers to form models. These models are applied on the test data for recognition. For classification process Bayesian network, Quest C5.0 and CHAID are considered. A brief description of the above classifiers is as follows

A. Bayesian Network

The Bayesian Network [11] enables to build a probability model by combining observed and recorded evidence with real-world knowledge to establish the likelihood of occurrences by using seemingly unlinked attributes. Bayesian Networks are mainly used to perform classification [12] task. They are based on conditional probabilities. Bayesian classifier learns from the training dataset. It supports for missing data during learning and classification. The algorithm is based on the conditional probability of each attribute. Let Xi be the attribute for a given class level Y. The algorithm applies Bayes rule to compute the probability of Y given instances of X1…Xn. It predicts the class with highest posterior probability. Some of the advantages of the Bayesian networks are backward reasoning, white box approach, allows rich structure and supports missing data.

B. QUEST (Quick, Unbiased and Efficient Statistical Tree)

Quest is a binary tree decision tree algorithm [13], [14]. The properties of Quest include unbiased variable selection, linear splits, imputation splits, handling categorical predictor variables and missing values. For ordinal and continuous attributes the algorithm computes ANOVA F–test or Levene's test. For nominal attributes it computes Pearson's chi–square (for nominal attributes). The spilt is based on the highest association between the input attributes and target attribute. Quest variable selection depends on the following steps

Let X1…Xn be the numeric predictor variables and Xl+1…Xk be the numeric predictor variables.

1) Find p-value from ANOVA F-test for each numerical variable.
2) Find p-value for each X^2-test for each categorical variable.
3) Choose variable $X_{k'}$, with overall smallest p-value $p_{k'}$

C. C5.0 Classifier

C5.0 is a decision tree algorithm used for classification [15]. It builds a decision tree or rule set by using training data based on the concept of information theory. The algorithm splits the dataset into one or other class at each node. The split is based on the maximum normalized information gain. Each subset generated from first split is further divided on a different field. The process is repeated until no more split is

found. Finally splits with low level are identified and the nodes which do not contribute significantly are pruned. C5.0 algorithm can predict categorical target. Some of the features of C5.0 are handling noise data, boosting, over fitting and reduced error pruning techniques. Steps involved in C5.0 algorithm are as follows

1) Check for base cases.
2) By using the training data construct decision tree.
   a) First find attributes with maximum information gain.
   b) Spilt the dataset into one or another class label.
   c) Repeat the process until no more spilt found.
3) For each sample apply the decision tree constructed for the target class.

D. CHAID(Chi-squared Automatic Interaction Detector)

CHAID stands for Chi-squared Automatic Interaction Detector [16]. It builds non-binary trees for the analysis of larger datasets. It relies on the Chi-square test to determine the best next split at each step. The steps involved in construction of decision tree are as follows

1) Create/ prepare categorical predictors.
2) Compute chi square test to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable
3) Select the spilt variable with the smallest adjusted p-value
4) Repeat the split process until no more spilt possible.

E. Ensemble model

Ensemble models [6], [7] are basically used to improve the classification accuracy of weak learners. In this paper the outcomes of weak learners are combined by using confidential weighted voting scheme to form a number of ensemble models. At first the outcomes of QUEST, Bayesian network and C5.0 are combined to form ensemble model1. The outcomes of ensemble1, CHAID, Bayesian network and C 5.0 are combined to form ensemble2 model. Further the outcome of ensemble2 model is further combined with C5.0 and Bayesian network to form ensemble3 model. Further ensemble4 model is combined with C5.0, Bayesian and CHAID to form ensemble4 model to increase the classification accuracy of the system.

VI. SIMULATIONS AND RESULTS

For simulation work the dataset is first normalized and features are extracted by using gradient based approach. Then PCA is applied to reduce the features from 2519 to 75. A number weak learners Bayesian network, Quest, C5.0 and CHAID are trained by using 3600 samples and tested on 400 samples. The outcomes of the classifiers are combined by using confidential weighted voting scheme to form a number of ensemble models. Further the output generated from ensemble model is combined with other models to generate new ensemble models. In this way a number of ensemble models are formed and their recognition accuracies are compared. It is found that every newly generated ensemble model is better than the previously generated ensemble models. Table 1, 2, 3 and 4 shows the confusion matrices for ensemble1, ensemble2, and ensembl3 and ensembl4 model respectively. Table 5 shows the classification accuracy of various individual and ensemble models. The performances of the models are also compared by using Gain and Response Chart [17]. Figure 6 and 7 represents the gain and response chart for all individual and ensemble model for target class '0'. From the charts ensembl4 model achieves highest recognition rate.

TABLE I. CONFUSION MATRIX FOR ENSEMBLE1 MODEL

	0	1	2	3	4	5	6	7	8	9
0	35	2	1	0	0	0	0	0	0	2
1	0	37	0	0	0	0	1	1	0	1
2	2	0	27	0	0	0	2	8	0	1
3	0	0	0	39	0	0	0	1	0	0
4	0	0	0	0	35	2	0	0	0	3
5	0	2	0	2	1	33	0	1	0	1
6	0	0	0	0	0	0	38	2	0	0
7	2	0	6	0	0	1	2	28	0	1
8	0	1	0	1	0	3	1	0	33	1
9	0	1	0	0	2	2	0	1	0	34

TABLE II. CONFUSION MATRIX FOR ENSEMBLE2 MODEL

	0	1	2	3	4	5	6	7	8	9
0	35	2	1	0	0	0	0	0	0	2
1	0	38	0	0	0	0	1	0	0	1
2	2	0	27	0	0	0	2	8	0	1
3	0	0	0	40	0	0	0	0	0	0
4	0	0	0	0	35	2	0	0	0	3
5	0	2	0	2	1	33	0	1	0	1
6	0	0	0	0	0	0	38	2	0	0
7	2	0	5	0	0	1	2	29	0	1
8	0	1	0	1	0	2	2	0	33	1
9	0	1	0	0	2	2	0	1	0	34

TABLE III. CONFUSION MATRIX FOR ENSEMBLE1 MODEL

	0	1	2	3	4	5	6	7	8	9
0	36	1	1	0	0	0	0	0	0	2
1	0	38	0	0	0	0	1	0	0	1
2	2	0	27	0	0	0	2	8	0	1
3	0	0	0	40	0	0	0	0	0	0
4	0	0	0	0	36	1	0	0	0	3
5	0	2	0	2	1	33	0	1	0	1
6	0	0	0	0	0	0	38	2	0	0
7	2	0	5	0	0	1	2	29	0	1
8	1	1	0	1	0	2	1	0	34	0
9	0	1	0	0	2	2	0	1	0	34

TABLE IV. CONFUSION MATRIX FOR ENSEMBLE4 MODEL

	0	1	2	3	4	5	6	7	8	9
0	36	1	1	0	0	0	0	0	0	2
1	0	38	0	0	0	0	1	0	0	1
2	2	0	27	0	0	0	2	8	0	1
3	0	0	0	40	0	0	0	0	0	0
4	0	0	0	0	36	1	0	0	0	3
5	0	2	0	2	1	33	0	1	0	1
6	0	0	0	0	0	0	38	2	0	0
7	2	0	4	0	0	1	2	30	0	1
8	1	1	0	1	0	2	1	0	34	0
9	0	1	0	0	1	2	0	1	0	35

TABLE V. CLASSIFICATION ACCURACY OF ALL INDIVIDUAL AND ENSEMBLE MODELS

Models	Bayesian	QUEST	C5.0	CHAID	Ensemble1	Ensemble2	Ensemble3	Ensemble4
Accuracy	67.25%	67.25%	79%	63.75%	84.75%	85.5%	86.25%	86.75%

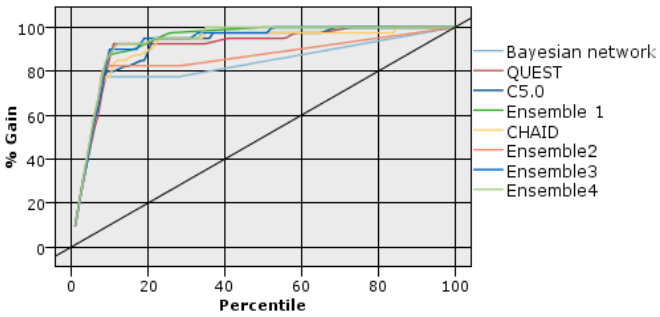

Figure 6 Gain chart for all individual and ensemble model for target class 0

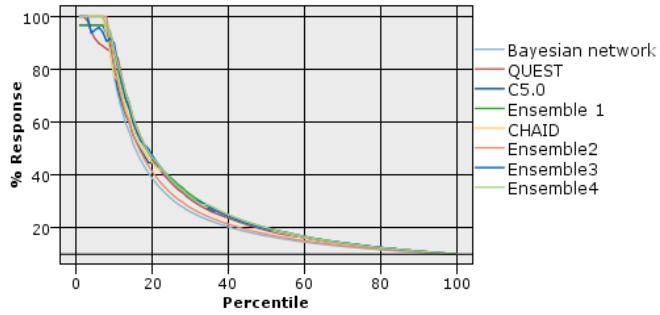

Figure7Response chart for all individual and ensemble model for target class0

VII. CONCLUSION

In this paper an attempt is made for recognition of Odia numeric digits by creating ensemble models with weak learners. Various character recognition steps preprocessing, feature extraction, feature reduction and classification steps are carried out. 4000 sample of Odia numeric digits are collected from ISI Calcutta. For classification 90% of the dataset is used for training and 10% of the dataset is used for training. The outcomes of the individual classifiers Bayesian Network, Quest, C5.0, and CHAID are combined by using confidential weighted voting scheme to form ensemble models. The ensemble models are further combined with other individual classifier to generate new ensemble models. The accuracy rate is found to be 67.25%, 67.25%, 79%, 63.75%, 84.75%, 85.5%, 86.26% and 86.75% for Bayesian network, QUEST, C5.0, CHAID ensemble1, ensemble2, ensembl3 and ensemble4 respectively. From the experimental result it is found that the ensemble4 yields highest accuracy with 86.75% as compared to other models. This exhibits the effectiveness of ensemble model over the weak learner. So the classification accuracy of the weak learners can be sufficiently increased by using ensemble model so as to increase the recognition rate of handwritten characters.

VIII. FEATURE RESEARCH

The proposed methodology can be further applied on online and offline optical characters. In this paper only handwritten odia numeric digits are considered. The proposed methodology can be applied on Odia handwritten scripts.

Other feature extraction methods apart from gradient based approach can be applied on the dataset. Besides the weak learner discussed in this paper other weak learners can be used with various combinations to produce more ensemble models so as to increase the classification accuracy of the system. The accuracy of the ensemble models can be tested on other well known datasets.

REFERENCES

[1] Mohamed Cheriet, Nawwaf Kharma, Ching Y. Suen," Character recognition Systems ", A guide for students and practioners, John Wielly and sons, Hoboken, New Jersey, 2007
[2] I.Czarnowski, P. Zedrzejowicz ,"Ensemble Classifier for Mining Data Streams, Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings, Volume 35, 2014, Pages 397–406.
[3] Muhammaad A. Tahir, James E Smith, "Feature Selection for Heterogeneous Ensembles of Nearest-neighbour Classifiers Using Hybrid Tabu Search", Advances in metaheuristics for Hard Optimization, Neural Computing Series, 2008, pp 69-85
[4] Koichiro Yamauchi,"Incremental Model Selection and Ensemble Prediction under Virtual Concept Drifting Environments," Lecture notes in computer science,PRICAI 2010, Trends in Artificial Intelligence,Volume 6230, 2010, pp 570-582
[5] P.Santos, L.F.Villa, A Renones, A. Bustillo, J. Maudes,"Wind Turbines Fault Diagnosis Using Ensemble Classifiers,Lecturer notes in Computer science Colume 7377,2012, pp 67-76 Advances in Data Mining, Applications and Theoritical Aspects.
[6] Z. Kamranian, S.A. Monadjemi and N. Nematbakhsh, "A novel free format persian/arabic handwritten zip code recognition system," Journal of Computers and Electrical Engineering, Vol. 39, pp:1970–1979, 2013.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[7] N. Das, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri and D.K. Basu, "A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application," Journal of Applied Soft Computing, Vol. 12, pp. 1592–1606,2012
[8] M. Shia, Y. Fujisawab, T. Wakabayashia, F. Kimuraa, "Handwritten numeral recognition using gradient and curvature of gray scale image," Journal of Pattern Recognition, Vol. 35, pp. 2051 – 2059, 2002
[9] B. Majhi, J. Satpathy and M. Rout, "Efficient recognition of odia numerals using low complexity neural classifier," In Proceedings of IEEE International Conference on Energy, Automation and Signal , pp. 140-143, 2011.
[10] S.Mitra, and T. Acharya, Data Mining: Multimedia, Soft computing, and Bioinformatics. New Jersy, Willy , 2004, pp. 204-205.
[11] Jiwaei Han, Kamber Micheline, Jian Pei Data mining: Concepts and Techniques, Morgam Kaufmann Publishers (Mar 2006).
[12] Machine Learning, 29,131-163(1997), Kluwer Academic Publishers, Neherlands.
[13] Lior Rpkach, "Data Mining with Decision Trees: Theory and Applications", World scientific , 2007
[14] http://www.math.ccu.edu.tw/~yshih/quest.html dated 15/02/14
[15] http://www01.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm dated 15/02/14
[16]]http://www.statsoft.com/Textbook/CHAID-Analysis dated15/02/14
[17] P. Pujari, "Classification and Comparative Study of Data Mining Classifiers with Feature selection on Binomial Data Set", Journal of Global Research in Computer Science", Volume 3, No. 5, May 2012,pp:39-45