2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) 978-1-7281-4889-2/20/$31.00 ©2020 IEEE 10.1109/ICCMC48092.2020.ICCMC-00016

Data Classification with k – fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques

Kaushika Pal
Sarvajanik College of Engineering & Technology
Surat, Gujarat, India
Kaushika.Pal@scet.ac.in

Dr. Biraj. V. Patel
G. H. Patel, P.G. Department of Computer Science & Technology, Sardar Patel University, V.V. Nagar, Gujarat, India
BirajPatel_4@yahoo.co.in

Abstract— Classification of documents is measured in terms of accuracy by comparing the actual labels with predicted labels for the classes. There are many machine learning techniques, which can be used to build a classifier, it is almost difficult to manually predict which technique should be used for classification, especially when we are working on Indic languages and there is no reliable method, which can give good results. Such area still need to be explored by processing documents with natural language processing, then applying machine learning techniques to build the classifier. This research article has collected 154 Hindi Poetries from web; processed it using NLP techniques, extracted features to build classifier with 5 Machine Learning methods namely support vector machine, naïve bayes, decision tree algorithm, random forest and k – nearest neighbors. The results were examined in accuracy with holdout accuracy estimator and k-fold cross validation accuracy estimator to check the reliability of the methods. The results of this research will help to select the best document classifier for Hindi Poetries and increase accuracy by working on various feature extraction techniques and NLP techniques. The result of the experiment shows that the results of SVM, NB and random forest methods are better as compared to DTT and K-NN for used data set available in this experiment.

Keywords—Mahine learning; NLP; classifier; accuracy; holdout; k-folds

I. INTRODUCTION

Hindi is the most common language of India and is very rapidly growing on World Wide Web. The content, which is growing exponentially, needs to be organized in order to be useful and easy access. Many researchers are working on many Indian languages to solve many issues related for easy retrieval and efficient storing of information. The current work is an effort to have organized Hindi data for efficient use. This experiment is working on documents containing written emotions of Romance, Courage and Sadness in form of Poems in Hindi language, which are named as Shringar, Karuna and Veera given for 3 different categories for Hindi poems in Sanskrit. The objective is to classify this documents into 3 classes mentioned and provide label to organize them for efficient retrieval.

The experiment conducted has processed 154 documents using natural language processing techniques; and features are extracted using machine-learning techniques. Machine learning algorithms namely Support Vector machine, Random Forest, K-Nearest Neighbor, Naïve Bayes and Decision tree are used for classification. To find the best supportive algorithms for Hindi Poetry classification is the goal of the experiment, to check the robust classifier it needs to be flooded with different data to see the variation in results; to achieve this k –fold cross validation is used. The outcome is best and robust classifier, which performs better on different variations of data set of Hindi Poetries for Multi-Class classification. Working on different poetic feature extraction can enhance accuracy of the classifier.

II. RELATED WORK

Jasleen Kaur, et al. [1] have experimented with 10 algorithms Adaboost, Bagging, C4.5, Hyperpipes, K- nearest neighbor, Naïve Bayes, PART, Support Vector Machine, Voting feature Interval and ZeroR to find the best algorithm for Panjabi poetry classification, they found HP, KNN, NB and SVM are better for Punjabi data. Harikrishnna D M, et al. [2] has classified Hindi Short stories with 3 algorithms K Nearest Neighbor, Naïve Bayes and Support Vector Machine to find the best one for classifying stories. Chaitanya Anne, et al. [3], classified patent document classification using kNN, SVM, J48, Random Forest and found SVM is working better for them. Mandal AK, et al. [4] has classified Bangla corpus using Naïve Bayes, Support Vector Machine, K - nearest Neighbor and Decision Tree and found SVM was performing better for Bangla data. S. Puri, et al. [5] proposed Hindi printed and handwritten Text Document Classification using SVM and Fuzzy logic, the text will be in image form; which will be processed and categorized into predefined classes. Rakhsit G, et al. [6] used multiclass SVM classifier to identify Poet of a poem belonging to 4 classes namely pooja, prem, prokriti, and swadesh, the data set was Bangla poetries. Senthil Kumar B, et al. [7] has surveyed on Feature Selection, Feature Reduction, and Machine Learning techniques for Text classification. Raj, Jennifer S. [8] has done survey on algorithms used for intelligent computing systems, Intelligent computing systems enables proper decision for any complex problems, the researcher discusses almost all data mining methods including clustering, regression, prediction, analysis, detection. Intelligent computing techniques discussed in the survey are recommended systems, supervised and unsupervised methods, which are branches of machine learning. Joseph, S. I. T. [9] has surveyed on tools of computational intelligence techniques and its applications, tools of computational intelligence are neural networks, fuzzy logic, genetic algorithms, belief networks, chaos and computational theory and artificial, which are used to build computation intelligent techniques along with computational model, the models are capable to process raw data and generate responses which are reliable and fault tolerant. K Pal, B.V.Patel [10] proposed model to classify poems in 9 categories using support vector machine and naïve bayes. The data set is Hindi Poetries. Shalini Puria [11] recommend a model for devanagari printed and handwritten character classification using SVM ensuring accuracy of 98.35% for handwritten characters and 99.54% for printed characters. Harikrishna D M, et al. [12, 13] extracted features to classify small stories using part of the speech and ESF, weighing schemes of keywords, part of the speech density, and analysis of part of the speech tags rendering to story genres. Chaitanya A, et al. [14], classified patent documents, using SVM, KNN, random forest and J48 machine learning algorithms.

Whenever there is new classification problem, there is no in built facility which helps you to find the model which will suit the dataset to be used, therefore extracting features in different ways and using different classification algorithms to find the best algorithm for the dataset is required. This research work is motivated and have an objective to select the best one out of 5 Machine Learning techniques which are widely used by experimenting with real data by dividing them into folds and and observing the results.

III. METHODOLOGY

The Method is divided into three basic modules

A. Pre-processing using natural language processing.

B. Feature extraction and numeric representation

C. Classification model

- training module
- testing module

Pre-Processing module is implemented in python 3.6 version, the module is tokenizing, removing special characters, symbols, numbers and stop words. Stops word list of 266 is manually created for the vocabulary used and was passed to the method removeStopwords() to remove it. The output of preprocessing module is files generated with only useful tokens. The files, generated are divided into k nearly equal portions as shown in fig.1 fold1 to fold n. n folds meaning n iterations of the model each time changing the train and test data set. Fig. 3 shows how actually the division of documents work.

In Feature Extraction module output files of preprocessing module is used to extract useful features, feature extraction is done on the basis of words occurring in documents and number of occurrence of each word ignoring the order of the word. The features are than converted to numbers and vector space model is created along with numeric representation of the feature and corresponding number of occurrence, which is bag-of-word model and this experiment is using unigram model.

Classification model comprises of 2 modules, they are training module and testing module

Training module will accept the output of feature extraction, which is feature set in numeric representation along with labels to train the model. This research article is using 5 machine-learning algorithms to train the model with the features extracted.

The support vector machine algorithm use hyperplane to categorize the documents, random forest creates forest of decision trees for classification; k-nearest neighbor finds the nearest neighbor to classify, decision tree classifies the documents by creating trees with nodes, it used gini index and entropy to decide the root and subsequent nodes to create tree. Naïve bayes use probability of documents to predict the class of new document.

All the techniques are very different from each other and therefore motivated to see at what accuracy the techniques are classifying the documents of Hindi poetries.

Testing module uses the trained model to predict the class of new Hindi poetries, which are not part of training.

The system used is shown in the fig 1. Where data is divided into n folds and fold 1 will be used for testing and rest folds will be used for training for iteration 1. Iteration 2 will take fold 2 for testing and fold 1, fold 3 to fold n for training and so on for all n folds.

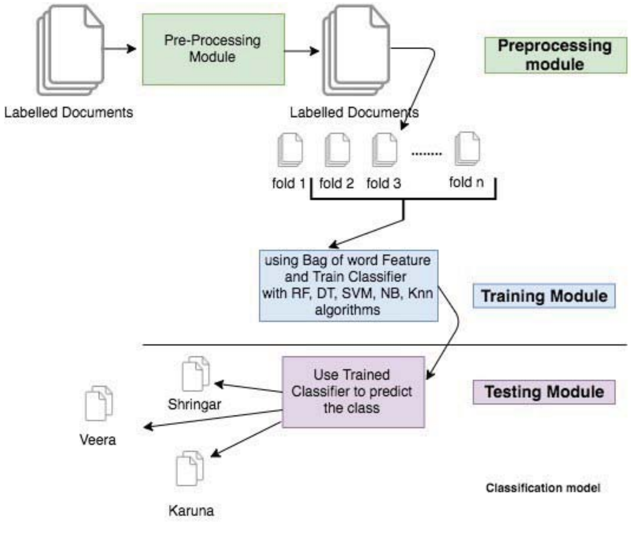

Fig. 1. Structure of the System

IV. Experiment and Performance Evaluation

The experiment uses 154 documents, which are collected manually from web and preprocessed, features were extracted and passed to the model for training with 5 different algorithms. The documents were divided into 2 portions of 80% and 20% for holdout accuracy estimates. The model was trained with 80% documents and tested with 20% documents. TABLE I shows numbers of poetries processed, words encountered and features extracted.

TABLE I. Statistics of Poems Processed

Poems Class	*No. of Poems*	*Total No. of words*	*Total Features extracted with Unigram*
Karuna	53	5116	1341
Shringar	52	5322	
Veera	49	8921	

The results of holdout accuracy estimator are shown in Table II and the visualized results are shown in fig.2.

TABLE II. Accuracy for Holdout estimator

Results	**Accuracy in Percentage**	
	Algorithm	***Accuracy***
RF	Random Forest	58.06
KNN	K Nearest Neighbor	32.25
CART	Decision Tree	45.16
NB	Naïve Bayes	51.61
SVM	Support Vector Machine	48.38

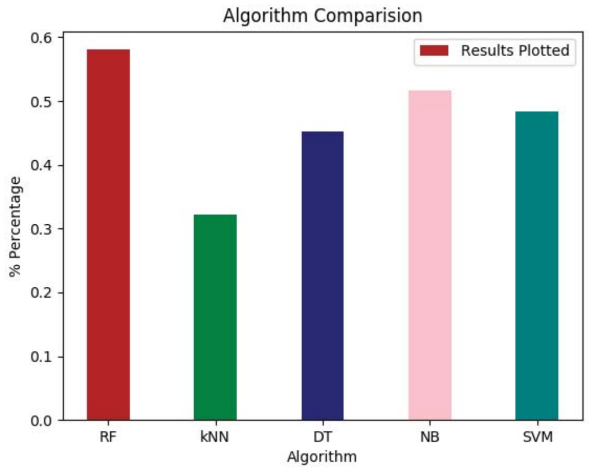

Fig. 2. Bar chart for Holdout accuracy estimator

K-fold cross validation divided documents into k portions and use 1 portion as test data and rest k-1 portion as train data. Fig. 3 shows the visualized version for k = 4 in fig [a] depicts that fold 1 will be used as test data and fold 2; fold 3; fold 4 will be training the model.

We have use k fold cross validation for k = 4, k = 6, k = 8, k = 10, k =12 and k = 14 for validating the results. The results of 4 folds are shown in Table III, and the boxplot which is showing clear range of results are shown in fig. 4 for k = 4.

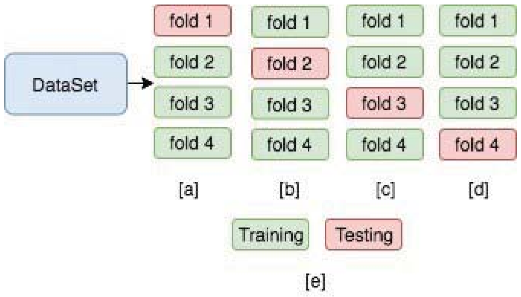

Fig. 3. K fold visualization for 4 folds

TABLE III. Accuracy for 4 folds

	Results in Accuracy of classification				
Algorithm/ Test Fold	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Average**
RF	51.42%	62.85%	55.88%	47.05%	54.30%
K – NN	37.14%	25.71%	54.28%	41.17%	39.57%
CART	37.14%	48.57%	42.85%	41.17%	42.45%
NB	62.85%	62.85%	45.71%	35.29%	51.68%
SVM	57.14%	62.85%	54.21%	44.11%	54.60%

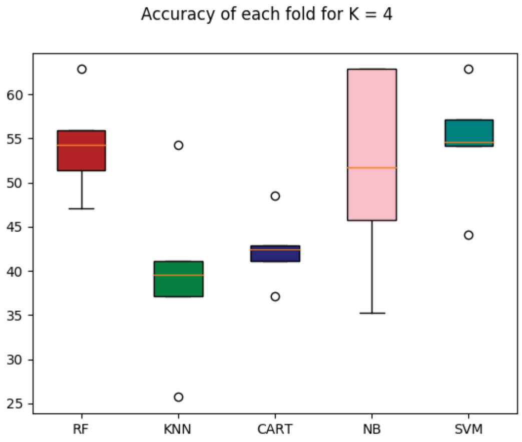

Fig. 4. Range in accuracy for each fold for K = 4

The results for k=6 was ranging from 34.78% to 58.33% for RF, 21.73% to 60.86% for KNN, 26.08% to 58.33% for CART, 30.43% to 65.21% for NB and 34.78% to 73.91%. The results are visualized in fig 5.

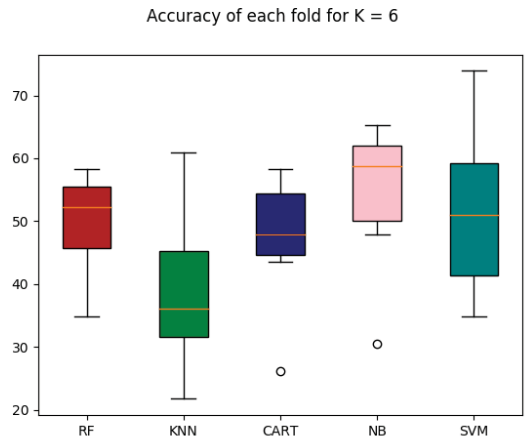

Fig. 5. Range in accuracy for each fold for K = 6

The results for k=10 was ranging from 28.57% to 64.28% for RF, 7% to 42.85% for KNN, 21.42% to 57.14% for CART, 30.76% to 64.28% for NB and 30.76% to 78.57%. The results are visualized in fig 6.

Similar results are predicted by the model for all the folds ranging from k=4 to k=14. The results for some folds of SVM are fantastic with 78.57% accuracy, which gives us a hint to explore more on SVM model with kernel 'linear' with more data variation.

The results also point to some outlier poetries, which need to be noticed and resolved for large variation in results for each fold.

The average results of each k folds ranging from 4 folds to 14 folds for all the algorithms used are shown in Table IV.

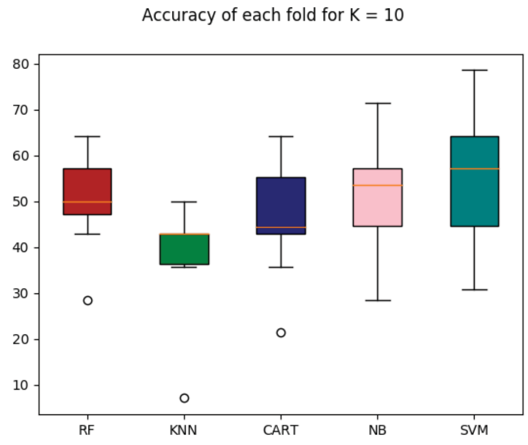

Fig. 6. Range in accuracy for each fold for K = 10

TABLE IV. AVERAGE ACCURACY FOR K FOLDS

	Average Results Accuracy in % for each k					
Classifier	**Avg. for K = 4**	**Avg. for K = 6**	**Avg. for K = 8**	**Avg. for K = 10**	**Avg. for K = 12**	**Avg. for K = 14**
RF	54.30	49.57	46.65	51.04	50.94	51.03
KNN	39.57	38.85	36.19	38.84	34.78	36.74
CART	42.45	46.67	47.50	46.04	37.81	45.23
NB	51.68	53.89	49.34	50.93	51.32	52.30
SVM	54.60	51.78	51.14	55.21	54.48	56.66

The results are interesting with accuracy for SVM for k-fold cross validation average results ranging from 51.14% to 56.66% with 'linear' kernel and 48.38% with holdout accuracy estimator, for Random Forest range is from 46.65% to 54.30% and holdout is 58.06%. Results of naïve bayes range from 49.34% to 53.89% and holdout accuracy is 51.16%. KNN is performing very poor with minimum accuracy of 34.78% and maximum of 39.57%, which is not acceptable for any classifier. CART is giving maximum accuracy of 47.50%.

The k – fold validation was used from 4 splits to 14 Splits. To show how close the average results for each K for each classifier is shown in fig. 7.

Results with k-fold cross validation shows that the results of random forest, naïve bayes and support vector machines are consistently higher as compared to KNN and CART algorithms. It is also observed that the tree model CART is not suitable for Hindi poetry classification but forest created using decision tree is giving good results.

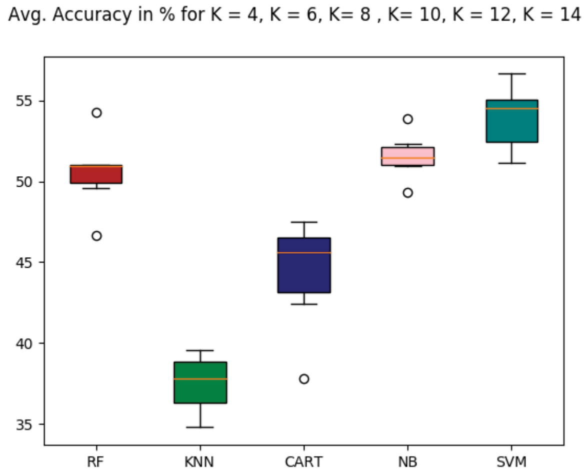

Fig. 7. Average Accuracy for each fold ranging from 4 to 14

V. CONCLUSION

To find the best classification algorithm to further enhance the performance, it is needed to see the results of each machine learning technique available for classification with different data sets. To provide different data set to model this research work has divided data into folds and providing data in different folds to the model, technically known as k-fold cross validation. Training with variety of data validates the robustness of the classifier to predict new data. This experiment has collected 154 poetries belonging to Shringar, Karuna and Veera class and tested 5 machine learning techniques with k fold cross validation performance estimator and found that RF, NB and SVM algorithms are better for Hindi poetries classification with given data set. The average accuracy range for RF is 46.65% to 54.30 %, NB is from 49.34 % to 53.89 % and SVM performance range is from 51.14 to 56.66. The results give us a direction to choose the best one and change the algorithm to enhance the results using chhand, alankar feature properties of Hindi poetries.

REFERENCES

[1] Kaur, Jasleen and Jatinderkumar R. Saini. “Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms.” International Conference on Machine Learning and Computing (ICMLC 2017)-ACM, *2017*

[2] Harikrishnna D M, K. Sreenivasa Rao. Children Story Classification based on Structure of the Story. IEEE International Conference on Advances in Computing, Communications and Informatics. (2015) 1485-1490

[3] Chaitanya Anne, Avdesh Mishra, Md Tamjidul Hoque, Shengru Tu, “Multiclass patent document classification”, Artificial Intelligence Research, Vol. 7, No. 1, 2017, pp. 1 - 14

[4] Mandal AK, Sen R. Supervised Learning Method for Bangla Web Document Categorization. International Journal of Artificial Intelligence and Applications. 2014; Volume 5 No. 5 pp. 93–105.

[5] Shalini Puri, Satya Prakash Singh. An Efficient Hindi Text Classification Model Using SVM Computing and Network Sustainability Book. 2019 .

[6] Rakhsit Geetanjali, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffari, “Automated Analysis of Bangla Poetry for Classification and Poet Identification”. In Proceeding of 12th International Conference on Natural Language Processing, Trivandrum, India, 2015, pp. 247 – 253.

[7] Senthil Kumar B, Bhabitha Varma E. A survey on Text Categorization. International Journal of Advance Research in Computer and Communication Engineering. Vol. 5, Issue8. 2016 pp. 286-289.

[8] Raj, Jennifer S. "A Comprehensive Survey on The Comutational Intelligence Techniques and it’s Applications”. Journal of ISMAC 1, no. 03. 2019 pp. 147-159

[9] Joseph, S. I. T. “Survey of Data Mining Algorithm’s for Intelligent Computing System”. Journal of trends in Computer Science and Smart technology (TCSST), 1(01), 2019 pp. 14-24.

[10] K Pal, B.V.Patel, “Model for Classification of Poems in Hindi Langauge Based on Ras”, Smart Systems and IoT: Innovations in Computing, Smart Innovation, Systems and Technologies 141. 2019 pp. 655-661.

[11] Duraipandian, M. "Performance Evaluation of Routing Algorithm for MANET based on the Machine Learning Techniques." *Journal of trends in Computer Science and Smart technology (TCSST)* 1, no. 01 (2019): 25-38.

[12] Harikrishnna, D.M., Sreenivasa Rao, K.: Classification of children stories in hindi using keywords and POS density. In: Proceedings of IEEE International Conference on Computer, Communication and Control. 2015

[13] Harikrishnna, D.M.K., Sreenivasa Rao, K.: Children story classification based on structure of the story. In: IEEE International Conference on Advances in Computing, Communications and Informatics,2015,1485–1490.

[14] Chaitanya Anne, Avdesh Mishra, Md Tamjidul Hoque, Shengru Tu, “Multiclass patent document classification”, Artificial Intelligence Research, Vol. 7, No. 1, 2017, pp. 1 - 14