International Journal of Engineering & Technology, 7 (4.19) (2018) 22-25

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparison of Algorithms in Authorship Identification using Bengali Poems

A.Pandian[1], V.Ramalingam[2], K.Manikandan[3], Payal Bhowmick[4], Shree Vaishnavi[5]

[1]Associate Professor, [2,3]Assistant Professor (S.G.), [4,5]B.Tech Student
[1,2,4,5] Dept. of CSE, [3]Dept. of Information Technology, SRMIST, Kattankulathur

Abstract

Author identification of Bengali poems is a paper mainly focusing on identification of an author of a poem. We train the system using a dataset consisting of features extracted from poems by various authors. Features like count of characters, words, spaces, vowels and consonants of Bengali poems are considered. Many training algorithms can be used to identify the authors. Some of the algorithms are J48, SVM, PCA, RDM, Random Forest Tree, Logic Regression, Naive Bayes etc. Every algorithm has its own advantages and disadvantages. The training algorithm used the most is J48 decision tree. It has additional features such as accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. which will be helpful when we want to classify with larger datasets.

Keywords: *Authorship identification; Bengali poems; J48 decision tree.*

1. Introduction

Author Identification is used in application of forensic analysis, humanities scholarship, electronic commerce, and in the development of computational methods for addressing various problems. Some authors write anonymously. It could be any kind of content; innocuous or detrimental. In some situations, to prevent harm, it is important to identify the author of a particular work. The data might be in any format like in text or as images. In earlier days people used to identify a document which was very time consuming and also expensive. Whereas the present electronic documents can be identified by using modern and automatic techniques in very less time. Certain properties are used to classify the poems which are called features.There are many ways to extract features like character count, word count, whitespace count, count of vowels and consonants of Bengali literature. These features are used to train a model using the proper algorithm until efficiency is achieved. The techniques of data mining are used for many purposes and tools such as Weka provide simple GUI that assist us to apply advanced machine learning algorithms to the datasets.

2. Materials and Methods

Rabindranath Tagore is the most globally renowned figure of Bengali literature. His notable Bengali work in poetry is Geetanjali, a book of poems for which he received the Nobel Prize for Literature. He acts as a bridge between earlier and later playwrights in terms of content and style.
Kazi Nazrul Islam wrote poems that light the fire against inequality or injustice and at the same time is known for his poignant romantic poems as well. Rajanikanta Sen, Atulprasad Sen, Dwijendralal Ray, Jatindramohan Bagchi, Kumud Ranjan Mullick, Jibanananda Das, along with Buddhadeva Bose, marks the beginning of the major move to transcend the Tagore legacy.Commonly called ”polli-kobi” (pastoral poet) Jasimuddin, Shamsur Rahman, widely known for his ’playing with words’ are also notable.
All of these poets in bengali literature have their own unique style that many other languages do not have.The vowels in Bengali called Sworoborno and the consonants are called Benjonborno. There are some special characters called Juktokhor and matra. Juktokhor are the alphabet which are the combination of any two letters in Bengali literature. We have used the alphabet and some of the statistical features in making this project.Other elaborate information about the feature extraction, datasets, creating database and the process of training are given below.

2.1. Collection of Poems by Various Authors

The poems we have used here are called datasets. We have manually collected the poems of different authors from different sites. The datasets are of 100 poems of each 5 authors which was collected from a popular website for Bengali poems.[15]

2.2. Feature Generation

Feature generation has been done by using the datasets.It involves an automatic extraction of features. All the features, that we have used in this project are listed below.

2.3. Creating Database with Generated Features

We use Microsoft Excel to create a database containing the various features that were generated in the previous step. Each set of features correspond to each of the five authors. Hence with this dataset we can train a model.

2.4. Training using Weka Tool

Weka contains machine learning algorithms that can be used for data processing, classification, clustering and other such tasks.

2.5. Different Algorithms

J48

Advantage:
In J48 one can predict the possible outcomes and change events schematically. One can easily and quickly modify a decision tree.

Disadvantage:
Sometimes the attributes will have only discrete values.

Naive Bayes

Advantage:
Very simple, easy and fast to implement.

Disadvantage:
It can make probabilistic predictions. It cannot perform regression.

SMO

Advantage:
It can hold large amount of data.

Disadvantage:
The SVM training process is slow especially for big data. One of the reasons of being slow is that it requires a solution of an extremely large QP optimization problem.

RDM

Advantage:
Can be the most efficient maintenance program. It lowers the cost by eliminating unnecessary equipment maintenance.

Disadvantage:
Savings potential is not readily seen by management.

2.6. Support Vector Machine (SVM)

Advantage:
Shows more accuracy when compared to other algorithms.

Disadvantage:
Takes long training time and not easy to incorporate domain knowledge.

PCA

Advantage:
The main advantage is due to its low noise sensitivity. It reduces the complexity in images by grouping with the use of PCA.

Disadvantage:
The covariance matrix is difficult to be evaluated in an accurate manner and also cannot capture a simplest invariance unless the training data provides information.

Hoeffding

Advantage:
It is a very fast decision tree algorithm and it is also a Concept adapting very fast decision tree algorithm.

Disadvantage:
Hoeffding tree algorithm wastes computational speed due to spending of lot of time in checking. It consumes more memory when expanding a tree.

2.7. Logistic Regression

Advantage:
It is more robust and also handles nonlinear effects.

Disadvantage:
It overstates the accuracy of its predictions.

2.8. Random Forest Tree

Advantage:
Runs efficiently on large databases. It can handle thousands of input variables without variable deletion.

Disadvantage:
It have been observed to over fit for some datasets with noisy classification or regression tasks.

2.9. Multilayer Perception

Advantage:
Multi-layer are most of the neural networks expect deep learning. it uses one or two hidden layers . The main advantage is they can be used for difficult to complex problems . However, they need long training time sometimes.

Disadvantage:
Multi layer perceptron with hidden layers have a non-convex loss function where there exists more than one local minimum. So initializing of different random weight can lead to different validation accuracy.It is sensitive to feature scaling.

2.10. LWL (Locally Weighted Learning)

Advantage:
LWL is also called lazy learning, because the processing of the training data is shifted until a query point needs to be answered. This approach makes LWL a very accurate function approximation method where it is easy to add new training points.

Disadvantage:
In LWL algorithm, sometimes no parameter values can provide a sufficient good approximation and also the computational costs are also very high.

3. Results and Discussion

In the comparison of 10 other algorithms in table 2, J48 algorithm has achieved an highest frequency of 87.4%. The Attribute selected classifier has also worked well by achieving 80.6% accuracy. The Multilayer Perception has achieved 65.6% accuracy and in the other hand Naive Bayes, Hoeffding and LWL has obtained an accuracy of 34 to 38%.

We have taken minimum number of objects and confidence factor as two important parameters.These two parameters helps us to obtain the best accuracy by using J48 algorithm. In table 3 we have taken the accuracies of 15 best features by taking a default confidence factor value of 0.25. The best accuracy we have obtained by performing this is 85.6% and in table 4 we have taken 2 as the default minimum number of object as we have obtained the highest frequency and the confidence factor is varied from 0.1 to 1. By performing this we have obtained the highest accuracy of 87.4% .

4. Conclusion

We can conclude that we have achieved a highest frequency of 87.4% accuracy by considering confidence factor as one of the parameter. The best accuracy we have achieved is by using J48 algorithm. In the paper[9], the dataset that they used are very small as compared to all the other papers. For future developments in author identification, there is scope of better accuracy using advanced algorithms.

Table 1: List of features

Features type	Features
Statistical:	
	1.Word (Count)
	Character (Count)
	Sentence/Line (Count)
	Average Word Length
	Paragraph (count)
	Whitespace (count)
	Mean of word (count)
	Median of word (count)
	Mode of word (count)
	Average Word length
	Average character in sentence
	Average character in paragraph
	Average words in sentence
	Average character in word
	Average words in paragraph
	Ratio of Whitespace to words
	Ratio of Whitespace to character
	Average Word Length
	Paragraph (count)
	Whitespace (count)
	Mean of word (count)
	Median of word (count)
	Mode of word (count)
	Average Word length
	Average character in sentence
	Average character in paragraph
	Average words in sentence
	Average character in word
	Average words in paragraph
	Ratio of Whitespace to words
	Ratio of Whitespace to character
Independent:	
	vowels count(Swarobomo count)
	consonants count(Benjonbomo count)
	special character count(Juktokhor count)
	matra (count)

Table 2: Accuracy of Classified Algorithms

S.no	Algorithm used	Accuracy Achieved
1	J48	87.4%
2	Attribute Selected Classifier	80.6%
3	Multilayer Perception	65.6%
4	BayesNet	47.2%
5	SMO	45.4%
6	OneR	42.8%
7	Naive Bayes	38.6%
8	Hoeffding	38.4%
9	LWL	34.6%
10	Naive Bayes Multinomial	33.2%

Table 3: Accuracies obtained while number of the best attributes considered range from 1 to 15

Attribute	Percentage
1	33.6 %
2	53.4 %
3	65.2 %
4	77.2 %
5	77.8 %
6	80.2 %
7	80.2 %
8	80 %
9	84 %
10	82.6 %
11	84.4 %
12	82.2 %
13	82.2 %
14	84.2 %
15	85.6 %

Table 4: Accuracies obtained while the minimum number of objects (MNO) range from 1 to 15

MNO	Percentage
1	96 %
2	85.6 %
3	78.4 %
4	73 %
5	69.4 %
6	66.8 %
7	66.4 %
8	63.8 %
9	63.8 %
10	62.2 %
11	60.6 %
12	60.6 %
13	57 %
14	57 %
15	56 %

Table 5: Accuracies obtained while the confidence factor Ranges from 0.1 to 1

Confidence factor change	Percentage
0.1	81 %
0.2	84.8 %
0.3	86.8 %
0.4	86.8 %
0.5	86.8 %
0.6	87.4 %
0.7	87.4 %
0.8	87.4 %
0.9	87.4 %
1	87.4 %

References

[1] Authorship Attribution in Bengali Language, Shanta Phani, Shibamouli Lahiri, Arindam Biswas, ltrc:iiit:ac:in=icon2015 =icon2015proceedings=PDF=37r p:pd f

[2] Automated Analysis of Bangla Poetry for Classification and Poet Identification using SVM classifier ,2015, Geetanjali Rakshit, Anupam Ghosh ,Pushpak Bhattacharyya, Gholamreza Haffari.

[3] Authorship Analysis and Identification Techniques: A Review , International Journal of Computer Applications (0975 – 8887), Mubin Shaukat Tamboli , Rajesh S. Prasad, Ph.D, Volume 77 – No.16, September 2013

[4] Author Identification in Bengali Literary Works using probabilistic classification method. , S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 220–226, 2011. Suprabhat Das and Pabitra Mitra ,Department of Computer Science and Engineering

[5] AUTHORSHIP ATTRIBUTION IN TAMIL CLASSICAL POEM (AGANANOORU): A MATHE-MATICAL MODEL, Dr.A.Pandian ,V.V.Ramalingam and R.P.Vishnu Preet , 2016.

[6] IDENTIFICATION OF AUTHORSHIP IN TAMIL CLASSICAL POEM (PARIPADAL) USING J48 ALGORITHM Dr.A.Pandian ,V.V.Ramalingam and R.P.Vishnu Preet , 2016.

[7] Author Identification based on Word Distribution in Word Space, 978-1-4799-8792-4/15/$31.00 c 2015 IEEE Barathi Ganesh H B*, Reshma U* and Anand Kumar M.

[8] Multi-Lingual Author Identification and Linguistic Feature Extraction — a Machine Learning Ap-proach, 978-1-4799-1535-4/13/$31 c 2013 IEEE , Hassan Alam, Aman Kumar.

[9] Author Identification for Digitized Paintings Collections, 978-1-4673-6143-9/13/$31.00 c 2013 IEEE, Razvan Condorovici, Corneliu Florea and Constantin Vertan

[10] Author Identification by Automatic Learning, 2015 13th International Conference on Document Anal-ysis and Recognition (ICDAR), 978-1-4799-1805-8/15/$31.00 c 2015 IEEE,Jordan Frery, Christine Largeron Laboratoire Hubert Curien

[11] Authorship Identification and Author Fuzzy ”Fingerprints”, 978-1-61284-968-3/11/$26.00 c 2011 IEEE, Nuno Homem , Joao Paulo Carvalho

[12] Author Identification using Sequential Minimal Optimization, 978-1-5090-2246-5/16/$31.00 c 2016 IEEE, John Jenkins, William Nick, Kaushik Roy, Albert Esterline, Joel Bloch

[13] Towards Author Identification of Arabic Text Articles, 2014 5th International Conference on Informa-tion and Communication Systems (ICICS), Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amez

[14] Author identification in Albanian language, 2011 International Conference on Network-Based Infor-mation Systems, Hakik PACI, Elinda Kajo, Evis Trandafili, Igli TAFA, Denisa Salillari.