Indian Journal of Science and Technology, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/107944, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using C4.5 Algorithm

A. Pandian*, V. V. Ramalingam and R. P. Vishnu Preet

Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India; pandian.a@ktr.srmuniv.ac.in, ramalingam.v@ktr.srmuniv.ac.in, vishnupreet.rp@gmail.com

Abstract

Objectives: To training classifier based on the features extracted from the poems of Mukkoodar Pallu, authors for various unknown poems can be classified. **Methods/Analysis:** The classification accuracy by performing classification in the dataset using C4.5 algorithm is illustrated in this paper. **Findings:** The results of performing classification on dataset that consists of features extracted from the dataset are shown in this paper. Features like number of characters, number of sentences and the classification accuracy when C4.5 algorithm is used is illustrated. **Novelty/Improvement:** By doing this, authors of various other poems in Tamil language can be identified which will be helpful to the society. Also a generalized authorship identification tool for all regional languages can be achieved.

Keywords: Authorship, Classification, Feature Selection, Tamil Articles

1. Introduction

Authors of many regional language poems are not yet identified. For instance, in Tamil language many poems are still anonymous. Identifying them would be of more use. Based on various researches, it turns out that most of the authorless poems can be associated with one of the authors, whose name and work is already known. So by using a suitable algorithm, authors for the unknown work can be identified. Thomas Bayes (1871) was the first to use statistical theory for solving authorship issues in the federalist papers. Auguste de Morgan as early as in 1851 has suggested the mean length of words as a measure to resolve authorship problem.

Identifying the writer of an article on the basis of stylistic character is the author attribution problem in linguistic research. Feature extraction can contribute more to this authorship problem, which consists of extraction of frequently used words, length of sentence, special characters used etc.

In[1], the authors explain how to extract features and find the accuracy of the classifier model. Using Enron dataset for e-mail and using 6 different algorithms, authors have achieved a maximum accuracy of 90.08%. 68.19% accuracy was achieved by using adaptive metropolis algorithm, 79.07% accuracy was achieved by using NBayes algorithm, 79.86% accuracy by using Bayes Net algorithm, 88.47% accuracy by using CMAR algorithm, 84.18% accuracy by using CBA algorithm and 90.08% accuracy by using CMARAA algorithm.

In[2], orders of components appropriate to Tamil utilizing bolster vector machine, proximal bolster vector machine and arbitrary kitchen sink calculations is performed. Bolster vector machine performs grouping by relegating focuses to one of the two disjoint spaces while Proximal Support Vector Machine arranges the dataset by allotting information focuses to the nearer of two parallel lines. Irregular Kitchen Sink calculation is a factual calculation that uses all the conceivable free factors. The exactness accomplished is 95.7%, 95.8% and 96.82% individually. In[3], demonstrate an exactness of 87.5% by utilizing irregular woodland calculation on 86052 words and 500788 characters.

In[4], 456 instances belonging to 7 authors of Arabic texts are used to perform classification using support vector machines, neural networks and markhov chains. Support Vector Machine performs classification by building a classifier model that assigns each example to

*Author for correspondence

either category, creating a non-probabilistic binary linear classifier. Neural networks are virtual abstraction of neuron cells that are present n human brain. These neural networks work the same way as neurons in brain are triggered. Markov chain algorithm performs classification only when the markov property is satisfied. A peak accuracy of 82% is achieved.

In[5], show the method to extract features from Tamil dataset that consists of 28420 characters and 5000 words with an accuracy of 72% to 82%. It uses FLD and RBF algorithms to overcome the overlapping problem. Fisher's Linear Discriminant algorithm performs classification by creating a linear combination of features that separates two or more classes of objects. Radial Basis Function algorithm is similar to artificial neural networks. It works based on the neuron parameters.

In[6], the author uses Arabic dataset to extract features from it and perform classification using markhov chain algorithm with an accuracy of 96.96%. The author explains clearly on how to extract features pertinent to Arabic and perform classification on it. All the features that are related to the Arabic dataset and that satisfy the markov property are only considered for classification. These features are selected and are used to build the classifier.

An exactness of 82% is accomplished on Arabic writings highlighted in[7], which utilizes bolster vector machine, neural system and markhov chain. The reference[8] demonstrates to concentrate highlights from old Tamil scripts that are digitalized, perform grouping on them utilizing bolster vector machine and bi-gram to achieve an exactness of 83%. N-grams are regularly gathered from discourse or content corpus. A n-gram of size one is called unigram and a n-gram of size two is called bi-gram.

In[9], the author solves the overlapping problem using fisher's linear discriminant and radial basis function algorithm by using Enron e-mail dataset, while in[10], the author explains how to extract features to find the authorship of an article by using radial basis function algorithm for classification in Enron e-mail dataset with an accuracy of 80% to 90%. The reference[11] shows how to recognize tamil letters from their ancient scripts by using Lab VIEW tool and performs classification on the dataset by using segmentation algorithm. The Enron email dataset was collected by CALO (Cognitive Assistant that Learns and Classifies), which consists of data collected from about 150 users.

The reference[12] contains a list of features that can be used to perform feature extraction from datasets. Classification was performed on Enron E-mail dataset using expectation-maximization and bisecting K-means algorithm that gives 90 % accuracy. In[13–15], authors explain various algorithms used to perform classification and their corresponding accuracy. The expectation-maximization algorithm is an iterative method to perform classification. This algorithm performs iteration between two steps E & M. The expectation step (E) creates a list of likelihood and the maximization step (M) expands the expected likelihoods listed in the expectation step.

2. Materials and Method

The present authorship identification methods support only English language. They do not support Tamil language. Finding the authors for unknown Tamil poems become difficult as there is no method to identify them. By extracting features pertinent to Tamil language and by using suitable algorithm, authors for these unknown poems can be identified. Classification is done by using text processing method. Text processing is the method of deriving high quality information from text that includes statistical patterns from the text.

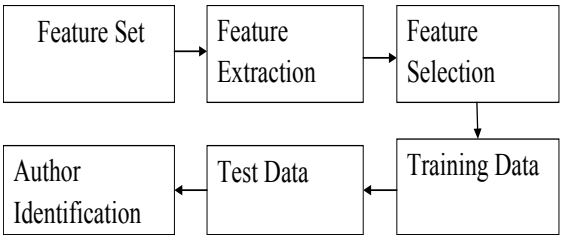

Figure 1. Architecture

The Figure 1 shows the architecture that is followed in this process of classification. The dataset considered here is "Mukkoodar Pallu" that consists of 800 instances anonymous poems. By extracting lexical, syntactic and semantic features as explained in the classification is performed. The list of features is shown in Table 1.

Table 1. List of Features

Features type Features
Lexical:
character-based
1. Character count (N)

2. Ratio of digits to N
3. Ratio of letters to N
4. Ratio of uppercase letters to N
5. Ratio of spaces to N
6. Ratio of tabs to N
7. Occurrences of uyir, mei and uyirmei letters (246 features)
8. Occurrences of special characters
Lexical:
word-based
9. Token count(T)
10. Average sentence length in terms of characters
11. Average token length
12. Ratio of characters in words to N
13. Ratio of short words (1 to 3 characters) to T
Syntactic
features
14. Occurrences of punctuations, . ? ! : ; ‘’”(8 features)

These features are extracted from the dataset and used for performing classification. These features define the stylometry of the author. Stylometry is the application of study of written styles from handwritten articles that can be used in authorship identification. Stylometry includes extraction of lexical, syntactic and semantic features pertinent to the language considered. Above table shows the lexical and syntactic features that are extracted from the dataset. By using the decision tree algorithm C4.5, an accuracy of 76.4% was attained.

Exactness of a classifier model is influenced by two parameters: Confidence variable and Number of elements considered. Certainty element is utilized to perform pruning of the choice tree. Progressively the estimation of certainty variable, additionally pruning will be finished. Least number of items alludes to the quantity of components to be considered. Both of these variables are utilized to perform choice tree pruning and maintaining a strategic distance from over fitting in the meantime. Grouping precision differs as the quantity of articles is shifted from 1 to the quantity of components considered. Weka device gives an approach to change these two parameters to enhance the classifier precision. By varying these two parameters, the accuracy of a classifier varies. Confidence factor can be varied from 0.1 to 1.0. The default value for confidence factor in Weka tool is 0.2. By varying this parameter, the classifier accuracy can be varied. Minimum number of objects (features) can be varied from 1 to number of features considered.

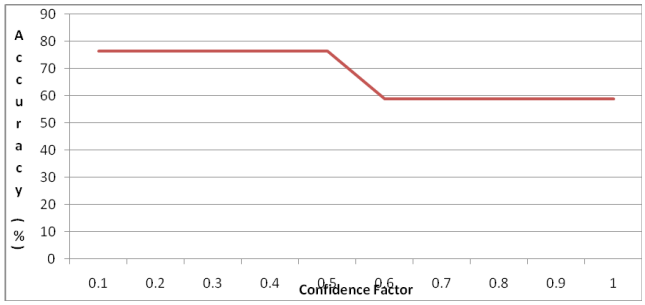

Figure 2. Confidence factor vs accuracy.

The graph (Figure 2) depicts the plot of confidence factor against classifier accuracy. It can be seen that the line falls down beyond confidence factor 0.5. From 0.1 to 0.4, the accuracy remains at 76%, so the default value 0.2 is fixed.

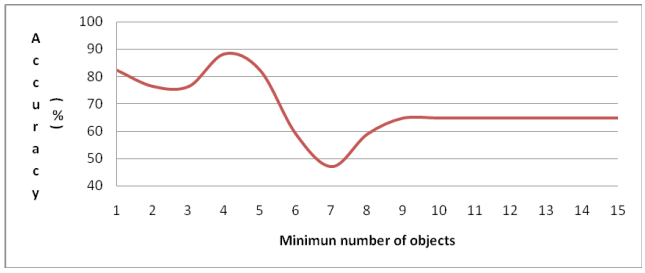

Figure 3. Minimum number of objects vs accuracy.

By fluctuating the quantity of items (components) from 1 to 15, changes in exactness can be seen (Figure 3). It can be seen that classifier exactness is at its pinnacle when the base number of items is 4. For different estimations of least number of items, the classifier precision is low. So the base number of items is settled as 4.

To achieve the most extreme classifier precision, the certainty variable is set as 0.2 and the base number of articles picked is 4. Out of 4 certainty calculate values, 0.2 is picked as it is the default esteem in weka apparatus that is utilized to play out the arrangement. Least number of articles is picked as 4 as it gives the most extreme classifier exactness.

2.1 Feature Extraction

Feature extraction process builds a set of derived values from the initial set of data that is intended to human interpretation. Dataset cannot be directly used in the

tool to perform classification. Only the features that are extracted from the dataset can be used to build the classifier.

Three types of features, lexical, syntactic and semantic are extracted. Lexical features include categories such as noun, verb, adjective, and pronoun. Syntactic features include noun phrase, verb phrase and prepositional phrase. Semantic features are those that include a set of features that intensifies the meaning of a word.

The features listed in Table 1 are extracted from the dataset. The dataset is first converted into Unicode formats so that it can be read in Microsoft excel. Computers cannot understand Tamil characters. They deal only with numbers in their memory. Unicode provides an encoding system that covers all the languages and provides a way for computers to understand them. UTF-8, UTF-16 and UCS-2 are the available Unicode encoding formats out of which UCS-2 is now obsolete. The encoding used in this process is UTF-16 which can be read in excel.

The extraction process is carried out by using macros, which can extract the specified features automatically. Macros are small programs that are used in Microsoft excel that can perform certain task repetitively to save time. The extracted features are in numeric format.

2.2 Feature Selection

Feature selection process is done by using decision tree. A decision tree is created using all the features that are listed in table-1 and the best features are selected based on the decision tree. The core algorithm to construct decision tree is ID3, which is now known as C4.5 algorithm.

Decision tree is constructed using two parameters: Entropy and Information Gain. The decision tree is constructed from root node and involves partitioning of nodes into subsets that consists of homogeneous objects. Entropy is used to measure the degree of homogeneity between the nodes that are present in a subset. Information gain increases as entropy decreases. Information gain and entropy are inversely proportional to each other. Decision tree construction is based on the attribute that contains the highest information gain.

Decision tree with all the features is pruned down to a number which provides maximum classification accuracy. Feature selection is done as it overcomes the problems of computational cost and inaccurate classifier accuracy due to irrelevant data. The features that are listed in table-1 are all selected by feature selection process as these features provide maximum classifier accuracy.

2.3 C4.5 Classification Algorithm

C4.5 algorithm is developed by Ross Quinlon. This algorithm is an extension of the ID3 algorithm that was in use earlier days. C4.5 algorithm constructs a decision tree from the set of training data that is used based on the entropy gain. This algorithm chooses each node based on the information gain, which is difference in entropy and splits its subsets effectively. The node with highest information gain or lowest entropy is used to make decision. This procedure is iterated for all the subsets until there are no further subsets to split. The steps of the algorithm is explained as follows:

1. Check for base cases.
2. For each attribute x, find the information gain by splitting on x.
3. Let x^1 be the attribute with highest information gain.
4. Create a node that splits on x^1.
5. Iterate on the subsets of x^1 and add all the nodes as children of x^1.

3. Results and Discussions

The confusion matrix obtained by performing classification using C4.5 algorithm is shown in Table 2. It can be seen that 11 instances of X are correctly classified while 2 instances of Y are incorrectly classified. The third attribute Z is classified correctly without any inaccuracies.

Table 2. Confusion matrix

	X	Y	Z
X	11	0	0
Y	2	0	0
Z	0	0	4

The parameters confidence factor and minimum number of objects are varied to improve the classifier accuracy. By default, the confidence factor in weka is 0.2. Figure 2 shows the graph of confidence factor against classifier accuracy. By varying the confidence factor from 0.1 to 0.4, the classifier accuracy obtained is 88.23%. By varying confidence factor from 0.5 to 1.0, the classifier accuracy obtained is 59%.

Similarly, other parameter minimum number of objects is varied from 1 to 14. This parameter is varied from 1 to 14 because the number of features considered is 14. Figure 3 shows the graph of minimum number of objects against classifier accuracy. By varying the number of parameters from 1 to 3, the classifier accuracy falls down from 83% to 78%. When the minimum number of objects is 4, maximum classifier accuracy of 88.23% is achieved. When the minimum number of objects is varied from 5 to 7, the classifier accuracy falls from 85% to 48%. Beyond 8, that is from 8 to 14, the classifier maintains a stable accuracy of 65%.

As the classifier extends a pinnacle exactness of 88.23% when the certainty variable is 0.2 and the base number of articles is 4, these two parameters are picked. From this, we can presume that the classifier gives an exactness of 88.23% when the certainty variable is 0.2 and least number of items is 4.

4. Conclusion

The features listed in table-1 were considered and the features were selected by constructing a decision tree using C4.5 algorithm. Certain features from the list of considered features list were selected in order to overcome over fitting of the classifier. The decision tree algorithm C4.5 produced an accuracy of 76.4% on the dataset. To improve the classifier accuracy, two parameters: confidence factor and minimum number of objects were varied. By choosing the confidence factor as 0.2 and minimum number of objects as 4, the classifier accuracy was increased to 88.23%. The authorship identification leads to an accuracy of 88.23% by varying these two parameters. Thus by extracting general features that are common for all regional languages, an overall authorship identification system can be developed for all regional languages.

5. References

1. Iqbal F, Binsalleeh H, Fung BCM, Debbabi M. Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation (Elsevier). 2010; 7:56–64.
2. Sanjanasri JP, Kumar MA. A computational framework for tamil document classification using random kitchen sink. IEEE, International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2015.
3. Khonji M, Iraqi Y, Jones A. An evaluation of authorship attribution using random forests. International Conference on Information and Communication Technology Research (ICTRC2015), IEEE; 2015.
4. Fawziotoom A, Abdullah EE, Jaafar S, Hamdellh A, Amer D. Towards author identification of Arabic text articles. 5th International Conference on Information and Communication Systems (ICICS), IEEE; 2014.
5. Pandian A. Sadiq MAK. Authorship categorization in email investigations using fisher's linear discriminate method with radial basis function. Journal of Computer Science. 2014; 10(6):10031214.
6. Ahmed A-F, Mohammad R, Bellahfkimustafa, Mohammad A-S. Authorship attribution in Arabic poetry. 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA); 2015.
7. Otoom AF, Abdullah EE, Jaafer S, Hamdallh A, Amer D. Towards author identification of Arabic text articles. IEEE, 5th International Conference on Information and Communication Systems (ICICS); 2014.
8. Urala KB, Ramakrishnan AG, Mohammad S. Recognition of open vocabulary. Online Tamil Handwritten Pages in Tamil Script. IEEE. 2014; 42(3):6–9.
9. Pandian A, Sadiq MAK. Detection of fraudulent emails by authorship extraction. International Journal of Computer Application. 2012; 41(7):7–12.
10. Pandian A, Sadiq MAK. Authorship attribution in Tamil language email for forensic analysis. International Review on Computers and Software. 2013; 8(12):2882–8.
11. Mahalakshmi M, Sharavanan M. Ancient Tamil script recognition and translation using Labview. IEEE, International Conference on Communication and Signal Processing; 2013 Apr 3–5.
12. Iqbal F, Binsalleeh H, Fung BCM, Debbabi M. E-mail authorship attribution using customized associative classification. Digital Vestigation; 2015. p. S116–26.
13. Bagavandas M, Hameed A, Manimannan G. Neural computation in authorship attribution: The case of selected Tamil articles. Journal Quantitative Linguistics. 2009; 16(2):115–31.
14. Chandrasekaran R, Manimannan G. Use of generalized regression neural network in authorship attribution. International Journal of Computer Applications. 2013; 62(4):7–10.
15. Pandian, A, Sadiq MAK. A study of authorship identification techniques in Tamil articles. International Journal of Software and Web Sciences. 2014; 7(1):105–8.